



eNTERFACE'19

The 15th International Summer Workshop on Multimodal Interfaces

PROCEEDINGS *eNTERFACE'19*

Summer Workshop
on Multimodal Interfaces

July 8 – August 2, 2019
Bilkent University – Turkey

Editors:
Hamdi Dibeklioglu, Elif Sürer

Proceedings eNTERFACE'19

15th International Summer Workshop
on Multimodal Interfaces

July 8 – August 2, 2019
Bilkent University, Turkey

Editors:
Hamdi Dibeklioglu, Elif Sürer

© Bilkent University, 2019

Published by
Computer Engineering Department
Bilkent University

Edited by
Hamdi Dibeklioglu
Bilkent University
06800 Ankara, Turkey
E-mail: dibeklioglu@cs.bilkent.edu.tr

Elif Sürer
Middle East Technical University
Graduate School of Informatics
06800 Ankara, Turkey
E-mail: elifs@metu.edu.tr

ISBN 978-605-9788-33-5

Cover Design
Hamdi Dibeklioglu

<http://enterface19.bilkent.edu.tr/>
<http://www.cs.bilkent.edu.tr/>



Bilkent University

Department of Computer Engineering

ORACLE®
Academy



Association for
Computing Machinery



Organization

General Chairs

Hamdi Dibeklioglu

Bilkent University

Elif Sürer

Middle East Technical University

Advisory Chair

H. Altay Güvenir

Bilkent University

Event Secretary

Ebru Ateş

Bilkent University

Technical Support Team

Berat Biçer

Bilkent University

Can Ufuk Ertenli

Middle East Technical University

Dersu Giritlioglu

Bilkent University

Burak Mandıra

Bilkent University

Vahid Naghashi

Bilkent University

Preface

The eNTERFACE workshops were initiated by the FP6 Network of Excellence SIMILAR. It was organized by Faculté Polytechnique de Mons (Belgium) in 2005, University of Zagreb (Croatia) in 2006, Boğaziçi University (Turkey) in 2007, CNRS-LIMSI (France) in 2008, University of Genova (Italy) in 2009, University of Amsterdam (The Netherlands) in 2010, University of West Bohemia (Czech Republic) in 2011, Metz Supélec (France) in 2012, New University of Lisbon (Portugal) in 2013, University of Basque Country (Spain) in 2014, University of Mons (Belgium) in 2015, University of Twente (Netherlands) in 2016, The Catholic University of Portugal (Portugal) in 2017, and Université catholique de Louvain (Belgium) in 2018.

The 15th Summer Workshop on Multimodal Interfaces eNTERFACE'19 was hosted by the Department of Computer Engineering of Bilkent University from July 8th to August 2nd, 2019. During those four weeks, a total number of 60 students/researchers from Europe came together at Bilkent University to work on seven selected projects on multimodal interfaces with diverse focuses including machine learning, virtual reality, video games, analysis of human behavior, and human-computer interaction. The titles of the selected projects were as follows:

- A Multimodal Behaviour Analysis Tool for Board Game Interventions with Children
- Cozmo4Resto: A Practical AI Application for Human-Robot Interaction
- Developing a Scenario-Based Video Game Generation Framework for Virtual Reality and Mixed Reality Environments
- Exploring Interfaces and Interactions for Graph-based Architectural Modelling in VR
- Spatio-temporal and Multimodal Analysis of Personality Traits
- Stress and Performance Related Multi-modal Data Collection, Feature Extraction and Classification in an Interview Setting
- Volleyball Action Modelling for Behaviour Analysis and Interactive Multi-modal Feedback

During the eNTERFACE'19 several excellent invited talks were delivered and we want to thank our invaluable speakers, (in order of appearance) Prof. Erol Şahin, Sena Aydoğan, Prof. Peter Robinson, and Prof. Albert Ali Salah, for their engaging and intriguing talks.

Those four weeks were not filled with research, projects, and keynote talks only; we also had a chance to visit the old town of Ankara (i.e., Ankara Castle and Rahmi M. Koç Museum) and enjoy the fairy-tale-like atmospheres of Cappadocia and Salt Lake together.

The organizers of eNTERFACE'19 would like to express their gratitude to the project leaders for their valuable proposals, and to all the participants and their funding institutions for their collaboration and excellent research outcome. After the intense research period enriched with social activities, all these projects achieved promising results, which are reported later in this document.

We want to thank our official sponsors Oracle Academy, Association for Computing Machinery (ACM), and Rahmi M. Koç Museum, for making this event possible.

We cannot thank enough Bilkent University, for hosting us during those four weeks, and the Department of Computer Engineering for their tremendous help in the organization. We thank our Advisory Chair Prof. H. Altay Güvenir for his generous support and dedication. We also want to thank our Event Secretary Ebru Ateş for her availability and responsiveness. A big thanks goes to our Technical Support Team (Berat Biçer, Can Ufuk Ertenli, Dersu Giritlioğlu, Burak Mandıra, and Vahid Naghashi) for their sincere help and contributions.

It was a great privilege to host you all in Ankara, Turkey while enhancing and enjoying this 15th edition of eNTERFACE together.



eNTERFACE participants after
the final presentations



eNTERFACE participants during
an invited talk



Trip to Salt Lake



Trip to Cappadocia

Program

Mon. July 8

- General opening meeting
- Project presentations
- Teams gathering and installation

Thu. July 11

- Invited talk: Erol Şahin (Middle East Technical University) – “The Notion of Affordance: Focusing on the Interface of the Agent with the World”

Sat. July 13

- Social event: Walking tour in the old town of Ankara

Tue. July 16

- Invited talk: Sena Aydoğan (Oracle) – “Oracle Academy: Resources for Education and Research”

Mon. July 22

- Invited talk: Peter Robinson (University of Cambridge) – “Computation of Emotions”

Tue. July 23

- Invited talk: Peter Robinson (University of Cambridge) – “Driving the Future”

Wed. July 24

- Midterm presentations
- Intermediate reports on teams achievements

Sat. July 27 – Sun. July 28

- Social event: Trip to Salt Lake and Cappadocia

Tue. July 30

- Invited talk: Albert Ali Salah (Utrecht University) – “Multimodal Analysis for Apparent Personality and Emotion Estimation”

Wed. July 31

- Gala dinner

Fri. Aug 2

- Final project presentations and concluding remarks

Table of Contents

MP-BGAAD: Multi-Person Board Game Affect Analysis Dataset.....	1
<i>Arjan Schimmel, Metehan Doyran, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander Bakkes, Heysem Kaya, Ronald Poppe, and Albert Ali Salah</i>	
Cozmo4Resto: A Practical AI Application for Human-Robot Interaction..	12
<i>Kevin El Haddad, Noé Tits, Ella Velner, and Hugo Bohy</i>	
Developing a Scenario-Based Video Game Generation Framework: Preliminary Results.....	19
<i>Elif Süner, Mustafa Erkayaoğlu, Zeynep Nur Öztürk, Furkan Yücel, Emin Alp Bıyık, Burak Altan, Büşra Şenderin, Zeliha Oğuz, Servet Gürer, and H. Şebnem Düzgün</i>	
Exploration of Interaction Techniques for Graph-based Modelling in Virtual Reality.....	26
<i>Adrien Coppens, Berat Biçer, Naz Yılmaz, and Serhat Aras</i>	
Spatiotemporal and Multimodal Analysis of Personality Traits.....	32
<i>Burak Mandıra, Dersu Giritlioğlu, Selim Fırat Yılmaz, Can Ufuk Ertenli, Berhan Faruk Akgür, Merve Kınıklıoğlu, Ashı Gül Kurt, Merve Nur Doğanlı, Emre Mutlu, Şeref Can Gürel, and Hamdi Dibeklioglu</i>	
Preliminary Results in Evaluating the Pleasantness of an Interviewing Candidate Based on Psychophysiological Signals.....	45
<i>Didem Gökçay, Fikret Arı, Bilgin Avenoğlu, Fatih İleri, Ekin Can Erkuş, Merve Balık, Anıl B. Delikaya, Atıl İlerialkan, and Hüseyin Hacıhabiboğlu</i>	
Volleyball Action Modelling for Behavior Analysis and Interactive Multi-modal Feedback.....	50
<i>Fahim A. Salim, Fasih Haider, Sena Büşra Yengeç Taşdemir, Vahid Naghashi, İzem Tengiz, Kübra Cengiz, Dees B. W. Postma, Robby van Delden, Dennis Reidsma, Saturnino Luz, and Bert-Jan van Beijnum</i>	

MP-BGAAD: Multi-Person Board Game Affect Analysis Dataset

Arjan Schimmel⁽¹⁾, Metehan Doyran⁽¹⁾, Pinar Baki⁽²⁾, Kübra Ergin⁽³⁾, Batıkan Türkmen⁽²⁾, Almıla Akdağ Salah⁽¹⁾, Sander Bakkes⁽¹⁾, Heysem Kaya⁽¹⁾, Ronald Poppe⁽¹⁾, Albert Ali Salah⁽¹⁾

⁽¹⁾ Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

⁽²⁾ Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

⁽³⁾ Sahibinden.com, Istanbul, Turkey

a.e.schimmel1@students.uu.nl, m.doyran@uu.nl, pinar.baki@boun.edu.tr, kubraergin3@gmail.com,
batikan.turkmen@boun.edu.tr, a.a.akdag@uu.nl, s.c.j.bakkes@uu.nl, hkaya@nku.edu.tr,
r.w.poppe@uu.nl, a.a.salah@uu.nl

Abstract—Board games are fertile grounds for the display of social signals, and they provide insights into psychological indicators. In this work, we introduce a new dataset collected from four-player board game sessions, recorded via multiple cameras. Recording four players at once provides a setting richer than dyadic interactions. Emotional moments are annotated for all game sessions. Additional data comes from personality and game experience questionnaires. We present a baseline for affect analysis and discuss some potential research questions for the analysis of social interactions and group dynamics during board games.

Index Terms—Board game, Dataset, Affect Analysis, Facial Modality, Social Interaction, Group Dynamics.

I. INTRODUCTION

Multiplayer board games are excellent tools to stimulate specific interactions both for children and adults. Many board games have been adopted for therapeutic purposes by psychologists that work with children [1], [2]. Players of board games exhibit a wealth of social signals. As such, they enable to study of affective responses to game events and other players and emotion contagion, possibly in interaction with personal and interpersonal factors. Board games have been used by therapists to assess behavioural patterns, a child's cognitive abilities, and attitudes [3], [4]. The assessments in turn may be employed for constructing playful interventions for children. Even though this approach is not a typical part of the toolkit of psychologists working with adults, a lot can be learned from analysing the game behaviour of adults as well.

Using board games for therapeutic purposes presents several methodological challenges. First, although a game may elicit valuable behavioural and affective responses, the amount of time when such a response can be observed during play is typically relatively brief [5]. Second, exhibited play behaviour (typically) cannot be easily annotated; accurate behavioural classifications not only depend on insight on human affect and decision-making processes, but also factors such as player personality and motivation, the state of the game, and the dynamics of the social context. Finally, manually coding a player's behaviour is inherently time-intensive. As such, while the potential for employing board games as analysis and

intervention tools is clear, at present it generally is time-consuming for therapists to fully exploit this potential.

With rapidly developing computational approaches to behavioural analysis, it is becoming increasingly more feasible to automatically process large amounts of play observations and, if needed, prepare indices for therapists. As such, an effective computational approach to behavioural analysis would mitigate the above-mentioned challenges. First, depending on the observed behaviour, a limited number of observations may suffice for accurate analysis. Second, multiple modalities such as the face, body and the voice can be analysed simultaneously; information from one modality may be used to assess the accuracy of classifications derived from other modalities. For example, knowing a person's head orientation may tell us something about the expected accuracy of facial expression analysis (*cf.* [6]). Third, automated analysis can be expected to be significantly more efficient than manual analysis. The drawbacks of fully automatic analysis are the limited generalization capabilities of such algorithms, their dependence on rich annotations (which may imply a small number of affective states as target variables, or in the case of continuous affect space annotations, a non-trivial mapping to practically useful labels), and the lack of a semantic grounding, which makes interpretation of rare events and idiosyncratic displays impossible. However, as the capabilities of the automatic analysis tools grow, they are expected to play larger roles in the toolbox of the analysts.

In this paper, we investigate approaches for automated multimodal behavioural analysis of adults interacting with each other while playing different types of board games. We introduce MP-BGAAD, a dataset with recordings of 62 game sessions, each involving four players. Using MP-BGAAD, we investigate to what extent we can derive information on personality traits, emotional states and social interactions of adults from recordings of their behaviour. Our setup includes the recording of videos of interacting players and the game board, the collection of personality traits for each player, and an assessment of the game experience after each played game.

This paper makes the following contributions:

- 1) We introduce a multi-person dataset with three levels of

annotations (segment-level, session-level, and user level, respectively) of recorded game sessions.

- 2) We present baseline evaluation results on this dataset by using state-of-the-art feature extraction and classification methods.
- 3) We analyse and discuss the effectiveness of the employed classifiers.

We proceed with a discussion of related works on affect analysis and datasets. Section III introduces our dataset, Multi-Person Board Game Affect Analysis Dataset (MP-BGAAD). We explain the board games, characteristics of the participants, recording setup, annotation process and the questionnaires we have used for assessing personality and game experience. In Section IV, the feature extraction process and classification methods are detailed. We present baseline scores for several automated analysis tasks in Section V and conclude with a discussion in Section VI.

II. RELATED WORK

Affective computing aims at equipping computer programs with the ability to sense affective cues exhibited by humans, in the hope of using these for the design of more responsive interactive applications [7]. However, such analysis can also be directly used, for example by psychologists who observe, describe, and quantify behaviours during long-term therapy. Since the type of features that can be automatically derived from human behaviour analysis is vast [8], [9], a comprehensive review is not included here. Rather, we focus on the automatic analysis of player behaviour during game play.

Since the setting we use involves board games, we focus on a scenario where multiple persons are sitting around a table to play a game with materials on it. The most interesting states during such a scenario involve responses to the game events, or to other players, such as frustration, anger, elation, boredom, excitement, disappointment, concentration, puzzlement, expectation, pride and shame. Of particular importance is the display of these emotions in children, as play scenarios are particularly suitable for them. The behaviours giving away these states mostly happen above the table, so the focus lies on the upper body. While fidgeting may also be indicative, putting a camera under the table is not a desired solution.

The face is regarded as the most expressive part of the body [10], and there are works specialised in processing faces of children during game play or other activities such as problem solving [11], [12]. The eyes are in particular shown to be good indicators of a person's engagement with an activity [13], [14]. The use of the bodily motions alone in affect recognition is less common than using facial expressions [15].

One of the challenges in affect analysis with a broad range of affective states to detect is the fact that each particular affective state, with the possible exception of happiness manifesting in a smile, happens rarely. Thus, these problems are typically severely unbalanced in terms of sample distributions, and it is very important to study them in natural conditions. Facial displays are by themselves difficult to fully catch these states automatically, as the face is also deformed via non-emotional speech. The use of a multimodal system can increase the

performance. Moreover, facial and bodily modalities are the most widely used signals for automatic analysis of interactions [16]. Filntsis *et al.* addressed affect recognition during child-robot interaction, and illustrated how the combination of face and bodily cues in a machine learning algorithm could yield better results than the use of a single modality [17]. A similar finding in the application domain of serious games was reported by Psaltis *et al.*, where decision level fusion was employed and the individual modalities were fused with the help of confidence levels [18].

The use of the body as a modality provides some challenges. For facial expressions, Ekman and Friesen introduced the Facial Action Coding System (FACS) [19], [20], which provides an objective way to describe facial movements of the face. However, there is no clear and unambiguous mapping from action units to expressions; there are only indicators for a number of expressions, some strongly correlated, and some not. For example, the upwards movement of lip corners, coded as AU12, is a good indicator of a smile. Yet it does not immediately tell us whether it is due to genuine enjoyment, or used as a social back-channel signal [21].

For the body, such a system does not exist. The body language associated with certain emotions is usually described by how body parts move, but it is much more idiosyncratic [22], [23].

How to represent emotions and affect is still up for debate [15]. In 1981, and Kleinginna created an overview of the definitions of emotion that existed until then [24]. This gave 92 different definitions. There has since been many works on affect and what it precisely is [25], [26]. A working definition is given by Desmet [27]: “*emotions are best treated as a multifaceted phenomenon consisting of the following components: behavioural reactions (e.g. approaching), expressive reactions (e.g. smiling), physiological reactions (e.g. heart pounding), and subjective feelings (e.g. feeling amused)*”. This definition agrees with our aims, as in this project, our ambition is to create a dataset where participants' subjective feelings during gameplay and their expressive reactions can be predicted.

Visual behaviour and affect analysis have been applied to gameplay contexts [28]. Action recognition methods are used by many researchers for analysing sports games such as tennis [29], basketball [30] and football [31]. There are game consoles (such as XBox) designed to have capabilities to analyse users through audiovisual cues, for instance for showing relevant ads to them, depending on their age or behaviour. Some researchers point out the need for emotional appraisal engines for video games in order to achieve human-like interaction between the players and the non-player characters [32], [33]. This can be achieved to a degree through visual analysis of faces through a camera [34]. Elsewhere, face and head gestures are combined with posture to recognise affective states of people playing serious games [35]. Some existing datasets provide researchers with audio, visual or audiovisual data to aid research on affective computing and social interaction analysis. The Tower Game Dataset [36], Static Multimodal Dyadic Behavior (MMDB) dataset [37], Mimicry database [38] and the PInSoRo database [39] are some of the important resources with which it is possible to

Name	Year	Modality	Subj.	Subj. per Session	Sessions	Annotations	Labels
The Tower Game Dataset [36]	2015	V + A	39	2	112	Manual, continuous	Eye gaze, body language, simultaneous movement, tempo similarity coordination and imitation are rated using a six-point Likert scale
Static MMDB Dataset [37]	2016	V + A	98	2	98	Manual, discrete	Actions are classified
Mimicry Database [38]	2011	V + A	40	2	54	Semi-automatic, discrete	- Behavioural expression labels (smile, head nod, head shake, body leaning away, body leaning forward) - Mimicry/ non mimicry labels - Conscious / unconscious
PInSoRo Dataset [39]	2018	V + A	120	1 or 2 with 1 robot	75	Manual, discrete	- Task engagement - Social engagement - Social attitude
MP-BGAAD	2019	V	58	4	62	Manual, discrete	Emotional moments are annotated based on shown expressions

TABLE I
RECENT GAME BEHAVIOUR DATASETS. V = VIDEO, A = AUDIO

study social interactions between two adults, or child-adult and child-robot interactions.

The Tower Game Dataset [36] consists of audio-visual recordings of two players and focuses on the joint attention and entertainment during a game. Annotation of the dataset has been done with Essential Social Interaction Predicates (ESIPs). The static MMDB dataset [37] focuses on dyadic interactions between adults and 15- to 30-month old children. The dataset is annotated based on the action-reaction dynamics. Multimodal Mimicry database [38] is recorded during two experiments: a discussion on a political topic and a role-playing game, respectively. The annotation consists of a number of social signaling cues and conscious/non-conscious labels illustrating the status of these cues. The PInSoRo dataset [39] has recordings of both child-child and child-robot interactions. This dataset is annotated using three different social interaction codes, which are task engagement, social engagement and social attitude, respectively. These databases are all based on structured, short, two-person video segments. In Multi-Person Board Game Affect Analysis Dataset, four participants of a board game are recorded simultaneously during each session, which affords for more complex interactions between the participants. Table I summarises available game behaviour datasets and their characteristics.

III. DATASET

In this section, we introduce the Multi-Person Board Game Affect Analysis Dataset (MP-BGAAD). MP-BGAAD is collected during the eNTERFACE 2019 Summer Workshop on Multimodal Interfaces¹. The dataset features participants playing cooperative (co-op) and competitive board games. Every game session consisted of four participants, recorded by two separate cameras, and an additional recording of the board game itself to allow for the detection of in-game events. Every participant filled in a HEXACO personality test [40] and after every game, they completed the in-game and social modules of

the Game Experience Questionnaire (GEQ) [41]. In total, there are 62 sessions recorded. The study received ethical approval from the Internal Review Board for Ethical Questions by the Scientific Ethical Committee of Boğaziçi University.

In the following subsections, we will describe the games, participants, recordings, annotations and questionnaires. All images are reproduced with explicit permission from the participants.

A. Games

According to [42], there are four categories of games for therapeutic use: communication games, problem-solving games, ego-enhancing games, and socialization games, respectively. We use two types of games in the construction of the MP-BGAAD: communication games and ego-enhancing games, respectively. In communication games, competition plays a smaller role, and inter-player communication is the key [43]. Ego-enhancing games on the other hand trigger stress, feelings of competition and challenge. This potentially leads to conflicts between game players, creating emotional states like frustration, disappointment, anger, but also relief, triumph, elation, etc.

Each session consisted of four participants that played one of six multiplayer games, see Table II. The game that was played was chosen by the participants. Before playing, the rules of the game were explained by the experimenters. We briefly describe each of the six games and the benefits of using such a game.

Magic Maze is the most played game in MP-BGAAD. It is a cooperative game where players work together to achieve a common goal. The players win by collectively managing four game characters exploring a maze. These characters need to steal certain items from specific locations of the maze, and use specific escape locations to complete the task against a running hourglass. Players do not take turns and are allowed to move whenever they can. Each player has a complementary set of moves. The game is played in real-time and if the hourglass (green circle in Figure 2) runs out, the players lose the game.

¹For more information about the workshop: <http://web3.bilkent.edu.tr/enterface19/>.



Fig. 1. A screenshot from the recording stream, where four players respond to a players mistake.



Fig. 2. A moment in a Magic Maze game, where the red cone was just placed in front of the player on the left, who is confused about his expected moves.

Players are not allowed to speak with each other during most of the game. The only way they can communicate is using a big red cone (red circle in Figure 2), which can be placed in front of another player to indicate that the other player needs to do something. In Magic Maze, players naturally show emotions due to the tension generated by the game. The time pressure prompts the players to perform well, as every wrong move will set back the group as a whole. The most stress-related emotions can be seen at moments when the hourglass is about to run out and players try to reverse it by visiting special squares in the maze. Another clear moment is during the use of the red cone. If players place it in front of another player, this is generally done with a lot of enthusiasm to prevent face loss. The player who gets it might show a number of emotions, mostly frustration or confusion (e.g. left player in Figure 2). A game of Magic Maze takes around 10-15 minutes.

Qwixx is a competitive game, primarily based on luck. The players throw dice every turn and take some of them to cross off numbers, based on certain restrictions, on their own sheet. At the end of the game, the player with the most crosses wins. Each action disables a number of future actions (e.g. crossing a number may disable crossing smaller or larger numbers of the same color for the rest of the game), thus

Type	Games	Sessions	Minutes	Participants
Cooperative	Magic Maze	39	405	156 (57)
	Pandemic	2	78	8 (4)
	The Mind	1	6	4 (4)
Competitive	Qwixx	10	203	40 (17)
	Kingdomino	8	140	32 (17)
	King of Tokyo	2	73	8 (5)

TABLE II
THE GAMES PLAYED IN MP-BGAAD. NUMBERS BETWEEN BRACKETS ARE UNIQUE PARTICIPANTS.

the game requires the players to calculate and take risks. The emotions that are shown during a Qwixx game are mostly moments of surprise, both happy and sad when players see the results of the dice throw. Another commonly occurring type of emotion is 'schadenfreude,' i.e., enjoyment of an other player's demise. When a player cannot cross something off, other players typically enjoy these moments.

Kingdomino is also a competitive game, where players compete to create the most valuable kingdom. Every turn, players take a piece of land to place it in their kingdoms. The pieces work just like domino stones and have similar placement restrictions. New pieces are revealed at the start of every turn. This typically evokes emotions such as positive and negative surprise (see Figure 3 for an example). A player's choices directly influence the other players, as the piece of land can only be chosen by one player. This creates moments of friction between the players. In Kingdomino, boredom sometimes occurs when a player takes a long time to think. Players also take the opportunity to talk to other players to convince them to take a certain piece. Those moments show negotiation skills and how players react to each other.

Pandemic is a cooperative game where players try to save the world from an epidemic. Players need to work together to keep the outbreaks of diseases under control, while at the same time finding the cures for these diseases. The game decides where the next outbreak is, based on a deck of cards which players need to draw from every turn. This creates a lot of tension in these moments, because depending on which card is drawn, the game can swing in favor of the players or it

Name	Year	Modality	Subj.	Subj. per Session	Sessions	Annotations	Labels
The Tower Game Dataset [36]	2015	V + A	39	2	112	Manual, continuous	Eye gaze, body language, simultaneous movement, tempo similarity coordination and imitation are rated using a six-point Likert scale
Static MMDB Dataset [37]	2016	V + A	98	2	98	Manual, discrete	Actions are classified
Mimicry Database [38]	2011	V + A	40	2	54	Semi-automatic, discrete	- Behavioural expression labels (smile, head nod, head shake, body leaning away, body leaning forward) - Mimicry/ non mimicry labels - Conscious / unconscious
PInSoRo Dataset [39]	2018	V + A	120	1 or 2 with 1 robot	75	Manual, discrete	- Task engagement - Social engagement - Social attitude
MP-BGAAD	2019	V	58	4	62	Manual, discrete	Emotional moments are annotated based on shown expressions

TABLE I
RECENT GAME BEHAVIOUR DATASETS. V = VIDEO, A = AUDIO

study social interactions between two adults, or child-adult and child-robot interactions.

The Tower Game Dataset [36] consists of audio-visual recordings of two players and focuses on the joint attention and entertainment during a game. Annotation of the dataset has been done with Essential Social Interaction Predicates (ESIPs). The static MMDB dataset [37] focuses on dyadic interactions between adults and 15- to 30-month old children. The dataset is annotated based on the action-reaction dynamics. Multimodal Mimicry database [38] is recorded during two experiments: a discussion on a political topic and a role-playing game, respectively. The annotation consists of a number of social signaling cues and conscious/non-conscious labels illustrating the status of these cues. The PInSoRo dataset [39] has recordings of both child-child and child-robot interactions. This dataset is annotated using three different social interaction codes, which are task engagement, social engagement and social attitude, respectively. These databases are all based on structured, short, two-person video segments. In Multi-Person Board Game Affect Analysis Dataset, four participants of a board game are recorded simultaneously during each session, which affords for more complex interactions between the participants. Table I summarises available game behaviour datasets and their characteristics.

III. DATASET

In this section, we introduce the Multi-Person Board Game Affect Analysis Dataset (MP-BGAAD). MP-BGAAD is collected during the eNTERFACE 2019 Summer Workshop on Multimodal Interfaces¹. The dataset features participants playing cooperative (co-op) and competitive board games. Every game session consisted of four participants, recorded by two separate cameras, and an additional recording of the board game itself to allow for the detection of in-game events. Every participant filled in a HEXACO personality test [40] and after every game, they completed the in-game and social modules of

the Game Experience Questionnaire (GEQ) [41]. In total, there are 62 sessions recorded. The study received ethical approval from the Internal Review Board for Ethical Questions by the Scientific Ethical Committee of Boğaziçi University.

In the following subsections, we will describe the games, participants, recordings, annotations and questionnaires. All images are reproduced with explicit permission from the participants.

A. Games

According to [42], there are four categories of games for therapeutic use: communication games, problem-solving games, ego-enhancing games, and socialization games, respectively. We use two types of games in the construction of the MP-BGAAD: communication games and ego-enhancing games, respectively. In communication games, competition plays a smaller role, and inter-player communication is the key [43]. Ego-enhancing games on the other hand trigger stress, feelings of competition and challenge. This potentially leads to conflicts between game players, creating emotional states like frustration, disappointment, anger, but also relief, triumph, elation, etc.

Each session consisted of four participants that played one of six multiplayer games, see Table II. The game that was played was chosen by the participants. Before playing, the rules of the game were explained by the experimenters. We briefly describe each of the six games and the benefits of using such a game.

Magic Maze is the most played game in MP-BGAAD. It is a cooperative game where players work together to achieve a common goal. The players win by collectively managing four game characters exploring a maze. These characters need to steal certain items from specific locations of the maze, and use specific escape locations to complete the task against a running hourglass. Players do not take turns and are allowed to move whenever they can. Each player has a complementary set of moves. The game is played in real-time and if the hourglass (green circle in Figure 2) runs out, the players lose the game.

¹For more information about the workshop: <http://web3.bilkent.edu.tr/enterface19/>.

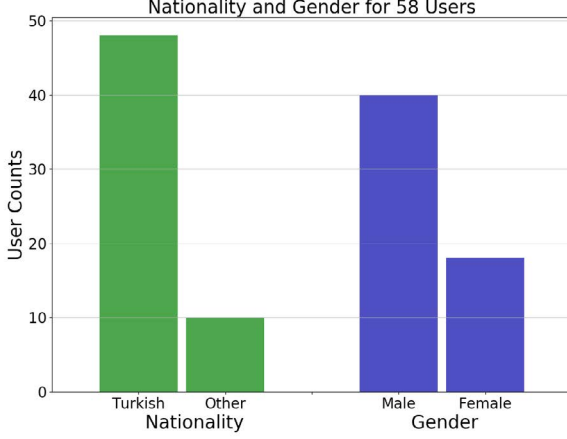


Fig. 6. Nationality and Gender histograms for all participants.

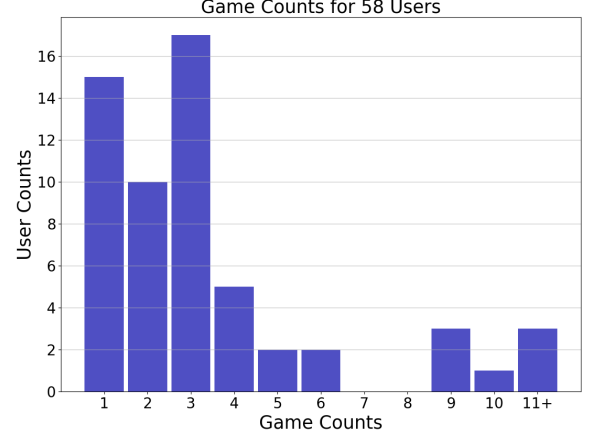


Fig. 8. Game count histogram for all participants.

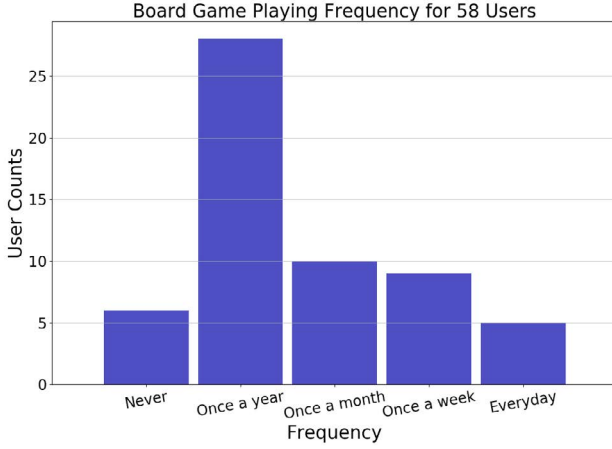


Fig. 7. Board Game playing frequency histogram for all participants.

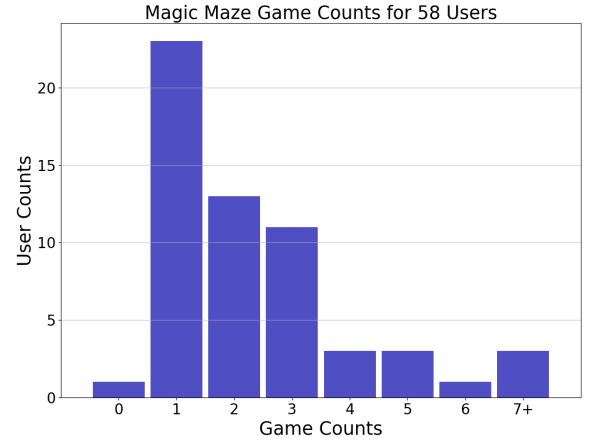


Fig. 9. Only Magic Maze game count histogram for all participants.

cameras are placed opposite to record both pairs (see left and middle frame in Figure 1). A third camera recorded the board and was placed slightly higher to have a better view (right frame in Figure 1). The three videos were merged into a single one (Figure 1) using Open Broadcaster Software² (OBS) to synchronize them for annotation purposes, but automatic processing is performed on the individual streams. The videos that were recorded of the participants (left and middle frame in Figure 1) have a resolution of 1280×720 and the recording of the board game state (right frame in Figure 1) had a resolution of 800×448 . All the recordings were captured in 30 fps. We decided not to focus on the audio in the recordings, because our recordings took place in a noisy environment. This would render the audio modality largely unsuitable. Furthermore, our participants were from different nationalities and they were not using their native language.

²<https://obsproject.com/>

D. Annotation

To mark expressive moments in the videos, we annotated for each player the deviations from a neutral facial expression. We used ELAN³ to create seven different annotations, Table III describes each in detail. People do not always show their emotions in the same way. For example, negative emotions can be expressed with a smile. If an anomaly shows but it was not clear what the label should be, the board game state was used to determine the label.

The dataset is annotated by two annotators with high inter-rater reliability. At the start of the annotating process, two videos were annotated by both annotators separately, and the final versions were compared. Annotators trained themselves further by discussing discrepancies in their annotations. After the training period, each video was annotated by a single annotator.

To measure the inter-rater reliability between the two annotators, we calculated Cohen's Kappa [44] on two videos that both annotators coded. Preliminary experiments have shown a

³<https://tla.mpi.nl/tools/tla-tools/elan/>

Label	Name	Meaning
+	Positive	Highest annotation in positive valence space. For example laughter and open mouth smiles.
+?	Small positive	Placed in positive valence space. Closer to the neutral state. For example gentle smiles.
'No label'	Neutral	Represents the state of the player that is generally shown throughout the game. Each player has a different neutral state, so annotations are done considering the most occurring state of that player.
-?	Small negative	Placed in negative valence space. Closer to the neutral state. For example short frowns and lowering of mouth corners.
-	Negative	Lowest annotation of negative valence space. For example looking angry and another player.
f	Focus	Not ranked in valence space. Player gives full attention to the board game. For example narrowed eyes and lower blink rate.
f?	Small focus	Less intense version of the focus label.
x	Non-Game event	For example taking a call or talking with another person outside of the game.

TABLE III
LABELS USED IN ANNOTATION.

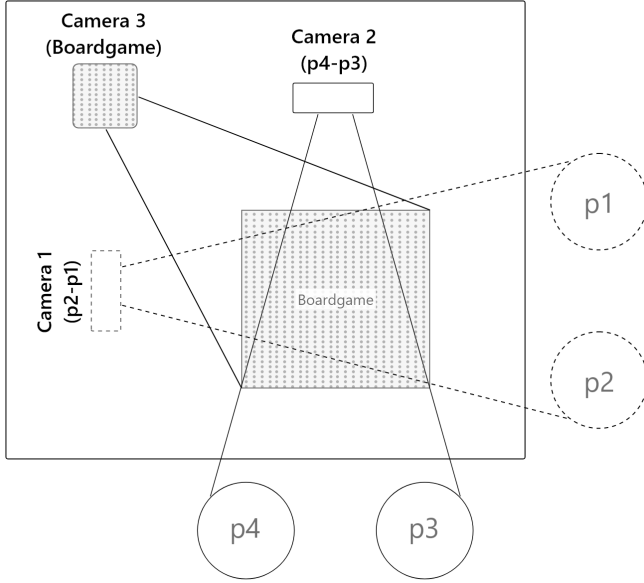


Fig. 10. Recording setup.

frame length of 50 to be adequate for segment-level coding. The Kappa score of our annotators was 0.735.

E. Questionnaires

The annotations of facial affect serve as in-game ground truth for the affective state of the player. To get social ground truths, the participants filled in two different questionnaires, which provided an opportunity to establish if there are correlations between the results of the recorded game data and the experience of the players reported by themselves. This also gave insights about the participants, and a baseline about checking if certain in-game events can be linked to certain personality traits.

Each participant filled in a 60-item HEXACO-PI-R test (HEXACO-60) [40] to assess personality in six dimensions: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience, respectively. Participants rated 60 statement sentences from 1 (strongly disagree) to 5 (strongly agree).

After playing a game session, each player filled in a Game Experience Questionnaire (GEQ [41]). The GEQ consists of

four separate modules, which can be used individually. We used the in-game and social presence modules to evaluate the participant's experience during the game, and to evaluate empathy, negative feelings and behavioural involvement with the other players, respectively. Players filled in the GEQ as many times as the number of game sessions they participated in. This gave MP-BGAAD 248 GEQ tests, which can be combined with the HEXACO-60 tests and in-game moments.

IV. METHODOLOGY

In this section and the next, we report some baseline approaches we have tested for the analysis of affect during board game play. We will first explain how the dataset is used to create features. Then, we will discuss how these features are used for automated detection of players' expressiveness.

A. Feature extraction

We have used the OpenFace 2.2 [45], [46]⁴ tool to locate faces in the video frames, and to extract facial landmark locations, head pose, eye gaze and facial expressions.

In each video, we have two players sitting side-by-side. As their relative positions do not change, tracking the nose landmark locations is sufficient to determine whether the output of OpenFace belongs to the left or right person in view. During a play session, it sometimes happens that a player reaches for something far away. The player then might appear in the recording of the other two players. To eliminate these unwanted faces that OpenFace detects, we check for clusters of outliers, corresponding to incidental face detections. To determine whether this process correctly labels the players with their assigned identities (from 1 to 4), we selected random frames and manually checked the outputs. From the selected frames, 100% was correctly labeled.

OpenFace provides a confidence score for each detection, which we used to exclude false or problematic detections from the feature set. The details of how this threshold affects the performance is presented in Table IV. Confidence thresholding gives us an improved feature set, but with missing frames and noise. To counteract these two problems we filter our feature set with a Savitzky-Golay smoothing filter [47]. We select this filter's window length (15) and polynomial order (3) empirically.

⁴<https://github.com/TadasBaltrusaitis/OpenFace>.

Models	Hyper-parameters			OpenFace confidence thresholds			
				0	0.25	0.5	0.75
Random Forest	with class weights	25 trees	depth:10	.5221	.5255	.5269	.5244
			depth:25	.3998	.4002	.4071	.3977
			depth:50	.3290	.3252	.3318	.3271
		50 trees	depth:10	.5250	.5278	.5285	.5271
			depth:25	.4020	.4021	.4075	.4040
			depth:50	.3132	.3126	.3139	.3114
		75 trees	depth:10	.5272	.5290	.5288	.5286
			depth:25	.4036	.4049	.4097	.4053
			depth:50	.3277	.3266	.3286	.3249
		100 trees	depth:10	.5276	.5290	.5293	.5299
			depth:25	.4031	.4066	.4102	.4040
			depth:50	.3199	.3180	.3193	.3165
	no class weights	25 trees	depth:10	.3584	.3610	.3615	.3682
			depth:25	.3847	.3886	.3914	.3888
			depth:50	.3814	.3939	.3875	.3942
		50 trees	depth:10	.3659	.3760	.3776	.3761
			depth:25	.3872	.3897	.3916	.3921
			depth:50	.3730	.3856	.3829	.3860
		75 trees	depth:10	.3693	.3825	.3840	.3801
			depth:25	.3992	.4023	.3997	.4034
			depth:50	.3887	.4014	.3988	.4018
		100 trees	depth:10	.3662	.3801	.3850	.3796
			depth:25	.3947	.3991	.3994	.4008
			depth:50	.3848	.3950	.3912	.3957
ELM	10 hidden units	rbf	tanh	.2730	.2713	.2720	.2682
			0.01	.0003	.0003	.0003	.0007
			0.1	.0003	.0003	.0003	.0007
	50 hidden units	rbf	tanh	.3681	.3730	.3730	.3715
			0.01	.3020	.2852	.2859	.2770
			0.1	.2849	.2727	.2749	.2698
	100 hidden units	rbf	tanh	.3737	.3686	.3705	.3719
			0.01	.3277	.3310	.3317	.3312
			0.1	.3287	.3304	.3322	.3307
K Nearest Neighbors	K = 3		.2779	.3021	.3054	.3110	
	K = 5		.2708	.2968	.2987	.3049	
	K = 9		.2542	.2782	.2825	.2867	
	K = 15		.2371	.2622	.2646	.2666	
	K = 31		.2166	.2377	.2396	.2430	
Decision Tree	with class weights	depth:5	.4986	.5087	.5089	.5063	
		depth:15	.4154	.4248	.4253	.4362	
		depth:30	.3501	.3428	.3500	.3533	
	no class weights	depth:5	.4092	.3980	.4082	.4092	
		depth:15	.3958	.3916	.3963	.3885	
			.3485	.3568	.3515	.3510	
Random			.2131				
All non-neutral			.2385				

TABLE IV
5 FOLD CROSS VALIDATION F1 SCORES ON THE TRAINING SET.

The processed data are then used to extract some features. These features are calculated over each small segment of a video, which are created with a sliding window approach. The window length (50 frames) and stride length (16 frames) are selected based on the best inter-rater agreement calculated in Section III-D. The features are divided into three categories: head movement (24), gaze movement (8) and action units (19), respectively.

- **Head movement:** OpenFace provides us with the location of the head in millimeters with respect to the camera. The location is given in 3D coordinates. We calculate the

Action Unit	Corresponding action
AU-04	Lowering of the brow.
AU-05	Raising of the upper eye lid.
AU-06	Raising of the cheeks.
AU-07	Tightening of the eye lid.
AU-09	Wrinkle in the nose.
AU-15	Lowering of the lip corner.
AU-20	Stretching of the lip.
AU-23	Tightening of the lips.
AU-26	Dropping of the jaw.

TABLE V
THE FACIAL ACTION UNITS USED IN THE ANALYSIS.

absolute movement of the head. The velocity and acceleration are calculated as the first and second derivative of the position with respect to time. OpenFace also provides the rotation of the head, in radians. These values can be seen as pitch, yaw, and roll. We calculate the absolute rotation to determine velocity and acceleration. For every segment, the mean and variance are calculated for the 3D coordinates of movement and pitch, yaw and roll for rotation. This provides us with 24 features for head movement.

- **Gaze movement:** OpenFace outputs the angle of the gaze by taking the average of the gaze vectors of both eyes. This creates two gaze angles in the horizontal and vertical direction. Similar to head movement, we calculated the mean and variance of the velocity and acceleration per segment. The result is eight features for the gaze.
- **Action Units:** OpenFace provides us with a subset of action units (AU), used to describe facial movements such as AU 45, which corresponds to the blink event, as well as an intensity value between 1 and 5. In the case of AU45, a value of 5 corresponds to a fully closed eye. Straightforward thresholding the smoothed intensity of AU45 as a function of time gives us the number of peaks per segment to determine the number of blinks. The other AUs that are used are shown in Table V. The mean and variance of the intensity are calculated for each AU.

B. Classifying Emotional Moments

We use the same sliding window segmentation used in feature extraction to match our frame-level annotations to the extracted features. Out of many classifiers available, we use Random Forests [48], Extreme Learning Machines (ELM) [49], K Nearest Neighbors, and Decision Trees [50] for our segment level classification task.

Although we have several class labels as explained in Section III-D, the distribution of the classes is extremely imbalanced. The neutral class dominates the others, with 86.05% of all the video segments labelled as neutral. Consequently, we combine the minority labels into a single class called ‘non-neutral’ for the baseline experiments. We perform binary classification to classify the non-neutral segments. Since our focus is correctly classifying the non-neutral segments to perform further analysis on them, we select the F1 score, which is harmonic mean of precision and recall, as our evaluation metric.

Models	Hyper parameters		OF conf. thresh.	F1	Precision	Recall
ELM	100 units	tanh	0.25	.42	.57	.33
K Nearest Neighbors	K = 3		0.75	.34	.46	.26
Decision Tree	class weights	depth:5	0.5	.52	.42	.68
Random Forest	class weights	100 trees depth:10	0.75	.54	.50	.60
Random				.24	.16	.50
All non-neutral				.28	.16	1.0

TABLE VI
TEST SET RESULTS.

V. EXPERIMENT AND RESULTS

Our experimental results, presented in this section, will serve as a baseline for future evaluations. We randomly split our dataset into 70% training set and 30% test set based on the game sessions, so that all the videos of any play session are only in one of the sets. That way we enable session-based group dynamics analysis such as social interactions and roles. Currently, we only use player based features and do not look into any social cues. We expect future research on this dataset to focus on extracting higher-level features.

Table IV shows our findings on the training set with 5-fold cross-validation comparing different hyper-parameters and OpenFace confidence thresholds. We selected the best hyper-parameters and OpenFace confidence threshold for each classifier to be used in the test set experiments. We present scores for two dummy generators in the last two lines of our tables for comparison. The first generator randomly guesses between neutral or non-neutral states, and the second one always classifies segments as non-neutral, our target class. Both in the training set and the test set experiments, the latter generator gets better F1 scores than the former.

The hyper-parameters we try in Table IV are class weights, tree counts, and maximum tree depths for Random Forest classifiers. We use class weights and maximum tree depth for Decision Trees as well. Class weights are selected inversely proportional to the class distribution in the training set. This gives higher priority to the minority class and the classifiers with class weights get the highest scores. ELM hyper-parameters are the hidden unit counts, activation functions and in the case of RBF, the kernel width.

The final results on the test set are presented in Table VI. As can be seen from the table, the best OpenFace confidence thresholds for all four classifiers are found to be different than zero. It means that some levels of thresholding cleans out the extracted data correctly. Unfortunately, ELM overfits to the neutral class and because of that its performance is lower than Random Forests and Decision Trees. ELMs are known to be affected by imbalanced data and require combination with proper data selection steps to overcome this issue [51]. In our experiments, we have not used any data selection steps and thus the different ELMs we used in our experiments performed

poorly compared to the other classifiers.

VI. CONCLUSION

We have introduced the Multi-Person Board Game Affect Analysis Dataset, MP-BGAAD, consisting of video recordings of players playing different types of board games engaging in multi-player interactions. Self-reported personality tests of all the players and the game experience questionnaires filled after every game session make this dataset open to many research directions.

We have presented some baseline scores for our frame-level affect annotations on the videos. Our test set experiments show that out of all four classifiers, the random forest with class weights to boost minority class predictions gets the highest baseline score, followed closely by a class weighted shallow decision tree. Our results show that state-of-the-art feature extraction tools and straight-forward machine learning techniques cannot get high accuracy results on our challenging dataset. We believe that these challenges will enable new research on the analysis of affect, social interaction, personality-game behaviour relationship, and game behaviour-game experience connection.

One of our future aims is extracting bodily motion features and to create multimodal classifiers. After that, we would like to create a new set of annotations which would facilitate research on social interactions and group dynamics.

ACKNOWLEDGMENT

The authors would like to thank Bilkent University for hosting eNTERFACE'19 and especially Hamdi Dibeklioglu and Elif Sürer for organizing it.

REFERENCES

- [1] E. G. Shapiro, S. J. Hughes, G. J. August, and M. L. Bloomquist, "Processing of emotional information in children with attention-deficit hyperactivity disorder," *Developmental Neuropsychology*, vol. 9, no. 3-4, pp. 207-224, 1993. [Online]. Available: <https://doi.org/10.1080/87565649309540553>
- [2] A. I. Matorin and J. R. McNamara, "Using board games in therapy with children," *International Journal of Play Therapy*, vol. 5, no. 2, p. 3, 1996.
- [3] D. Frey, "Recent research on selective exposure to information," in *Advances in experimental social psychology*. Elsevier, 1986, vol. 19, pp. 41-80.
- [4] E. T. Nickerson and K. B. O'Laughlin, "It's fun—but will it work?: The use of games as a therapeutic medium for children and adolescents," *Journal of Clinical Child Psychology*, vol. 9, 1980.
- [5] R. A. Gardner, *The psychotherapeutic techniques of Richard A. Gardner*. Creative Therapeutics Cresskill, NJ, 1986.
- [6] P. M. Blom, S. Bakkes, C. T. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers, "Towards personalised gaming via facial expression recognition," in *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.
- [7] R. W. Picard, *Affective computing*. MIT press, 2000.
- [8] A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, "Challenges of human behavior understanding," in *International Workshop on Human Behavior Understanding*. Springer, 2010, pp. 1-12.
- [9] A. A. Salah and T. Gevers, *Computer analysis of human behavior*. Springer, 2011.
- [10] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, 2018.
- [11] R. A. Khan, C. Arthur, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (liris-cse)," *arXiv preprint arXiv:1812.01555*, 2018.

- [12] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Face and Gesture 2011*. IEEE, 2011, pp. 30–35.
- [13] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE transactions on intelligent transportation systems*, vol. 4, no. 4, pp. 205–218, 2003.
- [14] J. Schwarz, C. C. Marais, T. Leyvand, S. E. Hudson, and J. Mankoff, "Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3443–3452.
- [15] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.
- [16] A. Schirmer and R. Adolphs, "Emotion perception from face, voice, and touch: comparisons and convergence," *Trends in Cognitive Sciences*, vol. 21, no. 3, pp. 216–228, 2017.
- [17] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction," *arXiv preprint arXiv:1901.01805*, 2019.
- [18] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2016, pp. 435–439.
- [19] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.
- [20] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on cd rom," *A Human Face*, Salt Lake City, pp. 77–254, 2002.
- [21] H. Dibeklioglu, A. A. Salah, and T. Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *European Conference on Computer Vision*. Springer, 2012, pp. 525–538.
- [22] C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, 2018.
- [23] I.-O. Stathopoulou and G. A. Tsihrintzis, "Emotion recognition from body movements and gestures," in *Intelligent Interactive Multimedia Systems and Services*. Springer, 2011, pp. 295–303.
- [24] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and emotion*, vol. 5, no. 4, pp. 345–379, 1981.
- [25] K. Mulligan and K. R. Scherer, "Toward a working definition of emotion," *Emotion Review*, vol. 4, no. 4, pp. 345–357, 2012.
- [26] E. Shouse, "Feeling, emotion, affect," *M/c journal*, vol. 8, no. 6, p. 26, 2005.
- [27] P. Desmet, "Measuring emotion: Development and application of an instrument to measure emotional responses to products," in *Funology*. Springer, 2003, pp. 111–123.
- [28] B. A. Schouten, R. Tieben, A. van de Ven, and D. W. Schouten, "Human behavior analysis in ambient gaming and playful interaction," in *Computer Analysis of Human Behavior*. Springer, 2011, pp. 387–403.
- [29] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06. New York, NY, USA: ACM, 2006, pp. 431–440.
- [30] M. Perše, M. Kristan, J. Perš, and S. Kovačič, "A template-based multi-player action recognition of the basketball game," in *In: Janez Pers, Derek R. Magee (eds.), Proceedings of the ECCV Workshop on Computer Vision Based Analysis in Sport Environments, Graz, Austria, 2006*, pp. 71–82.
- [31] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, "Football action recognition using hierarchical lstm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 99–107.
- [32] E. Hudlicka, "Affective game engines: Motivation and requirements," in *Proceedings of the 4th International Conference on Foundations of Digital Games*, ser. FDG '09. New York, NY, USA: ACM, 2009, pp. 299–306. [Online]. Available: <http://doi.acm.org/10.1145/1536513.1536565>
- [33] J. Broekens, E. Hudlicka, and R. Bidarra, *Emotional Appraisal Engines for Games*. Cham: Springer International Publishing, 2016, pp. 215–232. [Online]. Available: https://doi.org/10.1007/978-3-319-41316-7_13
- [34] A. Cruz, B. Bhanu, and N. Thakoor, "Facial emotion recognition in continuous video," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 1880–1883.
- [35] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 677–682. [Online]. Available: <http://doi.acm.org/10.1145/1101149.1101300>
- [36] D. A. Salter, A. Tamrakar, B. Siddiquie, M. R. Amer, A. Divakaran, B. Lande, and D. Mehri, "The tower game dataset: A multimodal dataset for analyzing social interaction predicates," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2015, pp. 656–662.
- [37] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421.
- [38] X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, and M. Pantic, "A multimodal database for mimicry analysis," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer Berlin Heidelberg, 2011, pp. 367–376.
- [39] S. Lemaignan, C. E. R. Edmunds, E. Senft, and T. Belpaeme, "The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics," *PLOS ONE*, vol. 13, no. 10, pp. 1–19, 10 2018. [Online]. Available: <https://doi.org/10.1371/journal.pone.0205999>
- [40] M. C. Ashton and K. Lee, "The hexaco-60: A short measure of the major dimensions of personality," *Journal of personality assessment*, vol. 91, no. 4, pp. 340–345, 2009.
- [41] K. Poels, Y. de Kort, and W. IJsselstein, *D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games*. Technische Universiteit Eindhoven, 2007.
- [42] C. E. Schaefer and S. Reid, "Game play," *New York: John Wiley and Sons*, 1986.
- [43] J. P. Zagal, J. Rick, and I. Hsi, "Collaborative games: Lessons learned from board games," *Simulation & Gaming*, vol. 37, no. 1, pp. 24–40, 2006.
- [44] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [45] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.
- [46] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
- [47] W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, pp. 669–672, 1990.
- [48] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [49] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 107–122, Jun 2011. [Online]. Available: <https://doi.org/10.1007/s13042-011-0019-y>
- [50] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [51] U. Mahdiyah, M. I. Irawan, and E. M. Imah, "Integrating data selection and extreme learning machine for imbalanced data," *Procedia Computer Science*, vol. 59, pp. 221 – 229, 2015, international Conference on Computer Science and Computational Intelligence (ICSCSI 2015).



Arjan Schimmel Received his B.Sc (2016) Computing Science, with a game technology track, and M.Sc (2019) Game and Media Technology from Utrecht University, the Netherlands. His research interests is in applying modern computing technologies, like machine learning, to human behavior and interactions.



Metehan Doyran is a PhD candidate and a Teaching Assistant at the Information and Computing Sciences Department of Utrecht University, the Netherlands. He works at Social and Affective Computing group under Prof. Albert Ali Salah's supervision. He received his B.Sc. (2015) and M.Sc. (2018) degrees in Computer Engineering from Boğaziçi University, Turkey. His B.Sc. final project was "Dynamic Role Allocation of Soccer Robots" and he reached with his team Cerberus to quarter finals in RoboCup SPL 2015 in China. His M.Sc. thesis was "Indoor Visual

Understanding with RGB-D Images Using Deep Neural Networks". Currently his main research interest is using artificial intelligence and computer vision techniques for human behavior analysis.



Pınar Baki was born in Trabzon, Turkey in 1995. She received her B.Sc. from the Computer Engineering Department of Boğaziçi University. She is currently doing her M.Sc. at the same department. For her masters thesis, she is trying to detect the mood states of bipolar disorder from multimodal data. She is also a research engineer at Arçelik, studying age, gender and emotion analysis from speech data.



Kübra Ergin was born in Turkey in 1993. She received the Industrial Product Design B.S. degree from Istanbul Technical University, Istanbul, Turkey, in 2018. After graduating, she started to work as a User Experience Designer at sahibinden.com, which is a classified web site, Istanbul, Turkey until 2019. Currently, she is working as a User Experience Designer in Accenture Industry X.0, Istanbul, Turkey. She is interested in cognitive science, cognitive design methods, digital design and digital interactive art installations. Her main purpose is to combine

machine learning methods with aesthetic production processes.



Batkan Türkmen received his M.Sc. degree in computer engineering from Boğaziçi University, Turkey in 2019 under the supervision of Prof. Albert Ali Salah. He obtained his B.Sc. degree in computer engineering from Bilkent University, Turkey in 2015. His research interests lie in the field of affective computing and deep learning to address human behavior analysis.



Almıla Akdağ Salah is a digital humanities scholar. Her research focuses on developing methodologies to close/distant reading of cultural objects with machine learning tools. Currently she works on two projects: one investigating the history of sex and body in the western culture, the second analyzing the effect of breathing on trauma. She is an associate professor of Industrial Design at Istanbul Şehir University, and works as an adjunct faculty at Utrecht University, Department of Information and Computing Sciences.



Sander Bakkes Sander Bakkes is an assistant professor at Utrecht University Dept. of Information and Computing Science and is affiliated with Utrecht Center for Game Research. He received his Ph.D. degree in artificial intelligence in video games. His research areas are adaptive interactive environments, game personalisation, applied gaming, automated game design, procedural content generation, player experience modelling, and artificial Intelligence.



Heysem Kaya completed his PhD thesis on computational paralinguistics and multimodal affective computing at Computer Engineering Department, Boğaziçi University in 2015. He has published over 40 papers in international journals and conference proceedings. His works won four Computational Paralinguistics Challenge (ComParE) Awards at INTERSPEECH conferences between 2014 and 2017; as well as three ChaLearn Challenge Awards on video-based personality trait recognition (ICPR 2016) and explainable computer vision & job candidate screening (CVPR 2017) competitions. His team was the first runner up in video-based emotion recognition in the wild challenge (EmotiW 2015 at ICMI). His research interests include mixture model selection, speech processing, computational paralinguistics, explainable machine learning and affective computing. He serves in editorial board of SPIRAS Proceedings, as well as reviewer in more than 20 journals including IEEE Trans. on Affective Computing, Neural Networks and Learning Systems, Multimedia, Image and Vision Computing, Computer Speech and Language, Neurocomputing, Speech Communication, Digital Signal Processing and IEEE Signal Processing Letters. He is a faculty member in the Social and Affective Computing Group of the Department of Information and Computing Sciences, Utrecht University.

2012 and 2013, he received the most cited paper award from Image and Vision Computing. In 2017, he received a TOP grant from the Dutch Science Foundation.



Ronald Poppe received a Ph.D. in Computer Science from the University of Twente, the Netherlands (2009). He was a visiting researcher at the Delft University of Technology, Stanford University and University of Lancaster. He is currently an assistant professor at the Information and Computing Sciences department of Utrecht University. His research interests include the analysis of human behavior from videos and other sensors, the understanding and modeling of human (communicative) behavior and the applications of both in real-life settings. In

2012 and 2013, he received the most cited paper award from Image and Vision Computing. In 2017, he received a TOP grant from the Dutch Science Foundation.



Albert Ali Salah is a Full Professor of Social and Affective Computing at Utrecht University, the Netherlands. He works on multimodal interfaces, pattern recognition, computer vision, and computer analysis of human behavior. He has over 150 publications in related areas, including the edited books Computer Analysis of Human Behavior (2011) and Guide to Mobile Data Analytics in Refugee Scenarios (2019). Albert has received the inaugural EBF European Biometrics Research Award (2006), Boğaziçi University Foundation's Award of Research Excellence (2014), and the BAGEP Award of the Science Academy (2016). He serves as a Steering Board member of eNTERFACE and ACM ICMI, as an associate editor of IEEE Trans. on Cognitive and Developmental Systems, IEEE Trans. Affective Computing, and Int. Journal on Human-Computer Studies. He is a Senior Member of IEEE, member of ACM, and senior research affiliate of Data-Pop Alliance.

search Excellence (2014), and the BAGEP Award of the Science Academy (2016). He serves as a Steering Board member of eNTERFACE and ACM ICMI, as an associate editor of IEEE Trans. on Cognitive and Developmental Systems, IEEE Trans. Affective Computing, and Int. Journal on Human-Computer Studies. He is a Senior Member of IEEE, member of ACM, and senior research affiliate of Data-Pop Alliance.

Cozmo4Resto: A Practical AI Application for Human-Robot Interaction

Kevin El Haddad ⁽¹⁾, Noé Tits ⁽¹⁾, Ella Velner ⁽²⁾, Hugo Bohy ⁽¹⁾

⁽¹⁾ Numediart Institute, University of Mons, Mons, Belgium

⁽²⁾ commercom, Amsterdam, The Netherlands

kevin.elhaddad@umons.ac.be, ellavelner@gmail.com, noe.tits@umons.ac.be,

hugo.bohy@student.umons.ac.be

Abstract—In this paper we report our first attempt on building a Human-Agent Interaction (HAI) open-source toolkit to build HAI applications. We present a human-robot interaction application using the Cozmo robot built using different modules. The scenario of this application involves getting the agent's attention by calling its name (Cozmo), then interacting with it by asking it for information concerning restaurant (e.g: "give me the nearest vegetarian restaurant"). We detail the implementation and evaluation of each module and indicate the future steps towards building the full open-source toolkit.

Index Terms—Human-Agent Interaction (HAI), Human-Robot Interaction, deep learning, Text-to-Speech Synthesis (TTS), Keyword Spotting, Automatic Speech Recognition (ASR), Dialog Management, Sound Localization, Signal Processing, Cozmo.

I. INTRODUCTION

THE past decades witnessed the rise of Human-Agent Interaction (HAI) systems such as conversational agents and intelligent assistants. This work aims at contributing to the improvement of HAI applications and their incorporation to our daily lives. HAI systems are generally formed of different modules with different task(s) each, communicating with each other.

We aim at building a toolkit containing such modules, as well as a framework with two main purposes:

- 1) controlling the agent's behavior in a user-defined way;
- 2) connecting these modules together in a single application so that they could be able to communicate with each other in a user defined logic;

The goal is to have a toolkit allowing the users the most freedom possible in the way they utilize it to build their HAI applications. The above mentioned modules would thus be usable either in an "off-the-shelf" mode (outside the framework) or in the framework defined here.

In the same perspective, in the future, modules will be incrementally added to this toolkit allowing a wider range of HAI applications implementations. Also, the framework is designed in a way to easily add and connect modules needed (toolkit's ones or user defined ones) in order to build customized HAI applications. This gives users more freedom on how to utilize the toolkit.

In order to evaluate the performance of the developed toolkit in building HAI systems, an application will be developed using it: Cozmo4Resto. This HAI application is an interaction

with the Cozmo robot ¹ during which Cozmo will give the user informations about restaurants based on the user's queries as described in further detail in Section III. This robot was chosen mainly because of the simplicity of integration in a python-based application (see also Section III).

Towards building this application, in this paper, we present the modules developed to be used for Cozmo4Resto and added to the toolkit, as well as the framework mentioned above. We will therefore first present the HAI-toolkit in general in Section II. Then detail the Cozmo4Resto application is explained further and the modules developed detailed in Section III. Finally the implementation of the platform for Cozmo4Resto is detailed in Section IV.

II. HAI OPEN-SOURCE TOOLKIT

We present here the first version of this toolkit ² that will be used to implement HAI application like Cozmo4Resto. It is implemented in a modular way and, as mentioned earlier, can be viewed either as a framework upon which modules are connected and the agent's behavior is controlled to build an HAI application or as a library of HAI-oriented modules usable outside the framework.

A. Modules

A module's task is to perform an action or a sequence of actions which is/are part of the agent's behavior and which is/are needed in the application implemented. The input-output of each module is implemented in an object oriented way and will have a specific and fixed format. This way, each module can be modified/replaced/improved without affecting the implementation of the others. This will help making the toolkit more generic.

B. Behavior Framework

The framework's main purpose is to allow the integration of all the different modules in a single HAI system. It can be summarized as a finite state machine [1] (FSM)-based system combined with a communication system.

The FSM is used to describe the agent's behavior. Each state corresponds to a specific behavior of the agent. In the

¹<https://anki.com/en-us/cozmo.html>

²<https://github.com/kelhad00/hai-toolkit>

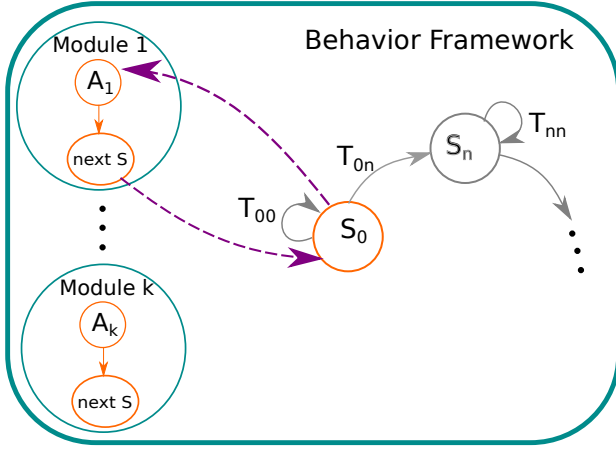


Fig. 1. Behavior framework workflow. S being the state, T the transition between states, A the action performed while in the state corresponding to the Module containing the action.

FSM, each state is linked to a module. Each module contains a sequence of actions the agent must perform while in the corresponding state.

The communication system allows the exchange of messages/data between the different modules. The main benefit of this is modularity and the ability to run each module on the same machine or different ones.

In what follows, we will refer to this framework as the "Behavior" framework, because it is used to describe the behavior of the agent through the states and the transitions between them. A visualization of the platform's workflow can be seen in Fig. 1.

Let $S = \{s_0, s_1, \dots, s_n\}$ be the set of states describing the agent's behavior (s_0 being the initial state), T_{ij} the transition from state i to j and A_k is the action performed by a module to which the state can be connected.

During run time, the Behavior framework acts as a client to each of the modules, which therefore act as servers. When in state s_i , Behavior queries the module to which s_i is linked with the input required. After the action is executed, a "next state" value is returned triggering the transition to another state or to the same current state.

The link between states and modules is defined by the user.

III. COZMO4RESTO APPLICATION

As mentioned above, to validate our toolkit, we use it to build an HAI application involving the Cozmo Robot.

Cozmo³ is a small physical relatively cheap robot which is designed to interact with users in games and other kinds of user-defined modalities. Cozmo can be very expressive through the eyes, audio and movements (body and head). A python SDK⁴ is provided with Cozmo for free, which makes it easy to intergrate on several platforms and with different applications. The SDK python commands are sent to Cozmo

via an android-based app that was developed for Cozmo. It contains a camera of which the stream is accessible via the SDK as well as other sensors. All this makes Cozmo an ideal platform to test our toolkit.

The interaction scenario of this application can be described as follows:

- 1) Cozmo would be wandering in a "non-interactive" mode;
- 2) when the keyword "Cozmo" is detected, Cozmo will turn around toward the caller and engage the interaction, thus going into "interactive" mode;
- 3) the user will query Cozmo concerning different information about restaurants (opening hours, menus, proximity, etc.);
- 4) after the interaction ends, Cozmo goes back to the "non-interactive" mode.

For this, three modules are needed: sound acquisition, keyword spotting (KWS) and sound localization (LOCAL). The sound acquisition module stores audio data in an efficient way for it to be used later on. The KWS detects a specific sound among others and LOCAL detects the directionality of the sound's source allowing to make Cozmo turn towards it. These will trigger the "interactive" mode. During the interaction an Automatic Speech Recognition (ASR) system will be used to convert the user's speech signal to text, which will be sent to a Dialog Management (DM) module. The DM module will take care of understanding the utterance and generating a text response based on an implemented logic (see Section III-E for a more detailed description of the dialog interaction). The response will be sent to a Text-To-Speech (TTS) synthesis system which will take care of converting the text response into an audio speech signal.

In what follows we will explain our approach to building and/or testing each module (some modules were already implemented open-source systems). The main constraints being the quality of the system and the computation time. Indeed a trade-off needs to be found so that the entire system generates high quality responses in a reasonable delay of time. We will also present the framework mentioned in Section II. The modules described in the following will be incorporated in the HAI-toolkit in general and are not meant only for the Cozmo4Resto application.

A. Sound Acquisition

The python pyaudio library⁵ is used to acquire the audio. A ring buffer is used to store and stream the input sound. The recording starts when the signal reaches a certain threshold. The recording stops when the signal goes below the threshold. The buffer is created by using the deque function from the collections python library. Each shift of audio signals recorded by the stream is added to the buffer, until it reached its maximum length. This maximum length is passed as a parameter at the creation of the buffer. Once the buffer is full, every new shift overwrites oldest one in the buffer. A number of channel (one per mic input) can be specified.

³Please note that at the time of writing this report, Anki, the company producing Cozmo, went bankrupted. But the SDK is still maintained at the time of redaction.

⁴<https://developer.anki.com/>

⁵<https://pypi.org/project/PyAudio/>

B. Sound Localization

Theory: The goal is to find the direction of arrival of a sound to a microphone or a set of microphones. For this we use the time delay of sound reception between the microphones. Three microphone positioned as shown in Fig. 2 and 3 are used here. The exact time of the sound emission from the source is unknown, so the time delays between reception time at each microphone are used instead. In Fig. 2, 'Source' is the sound's source and ' T_{mic_i} ' is the absolute time of reception of the microphone i . The source's coordinates (x ,

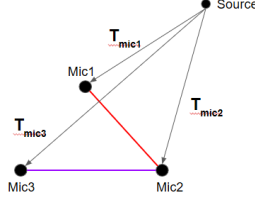


Fig. 2. Direction of arrival principle

y) are determined by minimizing the both following equations using the root() function from scipy.optimize python library :

$$v \cdot \tau_1 = \sqrt{(x_2 - x)^2 + (y_2 - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$

$$v \cdot \tau_2 = \sqrt{(x_3 - x)^2 + (y_3 - y)^2} - \sqrt{(x_2 - x)^2 + (y_2 - y)^2}$$

Where (x_i, y_i) are the microphone i coordinates and τ_i is the delay between $T_{mic_{i+1}}$ and T_{mic_i}

Technical setup: Two different hardware setups are considered for this module. The first one is composed of 3 AmazonBasics Microphones disposed in an equilateral triangular shape of side one meter (see Fig. 3). The second setup is composed of one Raspberry Pi 3 Model B and a 6-Mic Circular Array Kit from Sreed's Respeaker (see Fig. 4). Only 3 of the 6 microphones are used in the later. The red circles in Fig. 4 shows an example of the relative positions of the mics used among the 6 available for this setup.

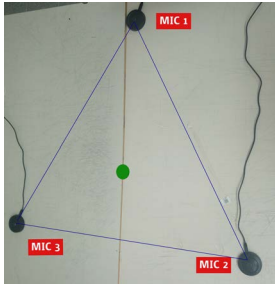


Fig. 3. Table Microphones setup

Database collection: In order to evaluate the above mentioned algorithm, we collected a dataset of different sounds with the different setups mentioned above. The sounds are either hand clapping or the word 'Cozmo', at different distances and angles with respect to the microphones: the angles vary by 30 degrees from 0 to 330 degrees and the distances are approximately 2m and 1m. The sounds were recorded in 2 different conditions:

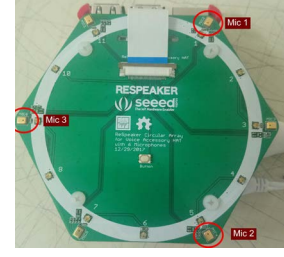


Fig. 4. Sreed's 6-Mic Array

- UMONS: recordings made in a relatively echo/reverberation-free room at 2m of the center of the microphones at the numediart institute of the University of Mons⁶.
- ENT: recordings made in a room generating echo/reverberation at distances $\geq 2m$ and $< 2m$ at Bilkent University⁷.

The algorithm described above was evaluated on the ENT condition only (data with reverberation) by calculating the cosine similarity between the angle estimated and the corresponding real value. The mean cosine similarity of all angles are shown in Table I in degrees, per sound and distance. This table shows the error between the actual position of the sound source and the estimated one.

TABLE I
MEAN COSINE SIMILARITY. COZ = COZMO, FAR=2M, CLOSE=1M,
SAME=AT TABLE HEIGHT, HIGH= 80CM HIGHER THAN TABLE HEIGHT

clap-far	clap-close	coz-far-high	coz-far-same	coz-close-same
26.19	10.89	6.77	6.04	2.82

The results indicate that the closer the source is from the microphones, the better is the estimation of the angle. These errors suggest that the type of sound might affect the efficiency of the sound localization algorithm used here. This is probably due to the difference in sound parameters like the sound amplitude and also the reverberation/echo generated by each sound.

Also, the effect this error might have on the user experience and user perception during an HAI application is an interesting and important aspect to consider.

Both of points will be investigated in future work.

C. Keyword Spotting (KWS)

As mentioned previously, the role of the KWS in this project is to trigger the "interactive" mode. In our case we use the keyword "Cozmo". A small dataset of "Cozmo" utterances from different speakers and in different tones was collected for the purpose of this work.

A benchmark of KWS systems is available comparing 3 systems online⁸: Picovoice, Snowboy and PocketSphinx. This benchmark uses crowd-sourced words to train and evaluate

⁶<https://numediart.org/>

⁷<https://w3.bilkent.edu.tr/bilkent/>

⁸<https://github.com/Picovoice/wakeword-benchmark>

TABLE II
AVERAGE WORD ERROR RATE (WER) AND DURATION OF COMPUTATION OF SENTENCES OF IEMOCAP DATASET

	Google Speech Recognition	DeepSpeech	Sphinx
WER	0.30	0.38	0.55
duration	1.69	0.8	9.5

these systems. The data used, therefore comes from different recording environments.

The customisation of Picovoice is done with text data. It relies on a dictionary of words with their corresponding phonetics. This dictionary is not accessible, and the word "Cozmo" needed for our application is not included in it. It is therefore not adequate for our application.

PocketSphinx is a mobile device version of Sphinx that runs locally, a group of speech recognition systems developed by Carnegie Mellon University. It uses HMMs for statistical modeling and includes a keyword spotting module. Similarly to Picovoice, it relies on a dictionary of words with phonetics.

Snowboy uses an API to send trigger words samples to train a system which is then downloaded and run locally. No more than three samples can be used to train the models.

Snowboy seems therefore like the best option for our application. We will therefore use "Cozmo" as training sample.

It is important to note that neither Snowboy or Picovoice systems are fully open source. Indeed the model training of Snowboy is performed through their web interface and Picovoice is optimized with a binary files provided online.

D. Automatic Speech Recognition (ASR)

Several ASR APIs are available for use under certain conditions. Some of these APIs are free but come with limitations of use in terms of API calls. Using APIs means we are dependent on a tier service that may not be free or not supported in the future.

A benchmark of APIs was proposed in [2]. Their code is open source⁹.

Research projects with open-source codes include Sphinx, DeepSpeech, Kaldi toolkit and gentle (based on Kaldi toolkit).

In this section, we compare APIs and State-of-the-art open source systems for our use-case. For this, we estimate their performance in terms of the Word Error Rate based on the Levenshtein distance. We use the IEMOCAP database to approach a setup closer to our use-case of interaction compared to a database based on Audiobooks recordings used in [2].

DeepSpeech may be accelerated with GPU. This was used on Google Colab with a Tesla K80 GPU.

Table II reports the average WER and duration for obtaining the prediction of the sentences of IEMOCAP dataset.

E. Dialog Management (DM)

To manage the dialogue between the agent and the user, a module needed to be implemented to extract the goal of the user's utterance and to give the right response back. This

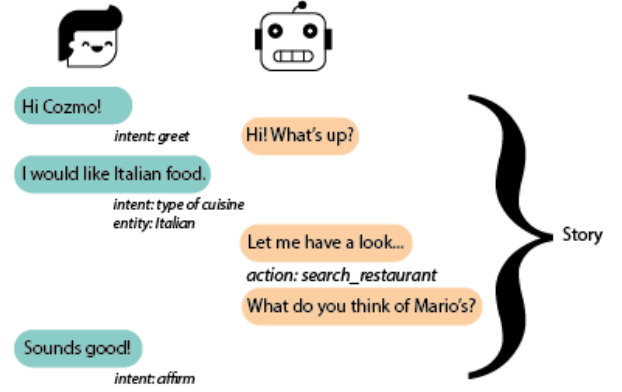


Fig. 5. Chat example with Rasa system.

is called a dialog management system. There are several options to create a working dialog management system, like using DialogFlow (Google) or Luis (Microsoft). However, we wanted an open-source library that worked with Python and JSON. Therefore, we chose the Rasa Library¹⁰.

RASA: This library is made up of two parts: Rasa NLU and Rasa Core. Rasa NLU extracts so-called 'intents' from the user's utterance. These intents are what the user wants or needs. These can be narrowed down by specifying 'entities'. For example, when the user says: "I would like Italian food", the intent is *the type of cuisine* and the entity is *Italian*. These intents and entities are then specified in a domain file, together with the template sentences that the agent can use to respond. This is where Rasa Core comes in. Rasa Core takes care of what the agent should do next, which most of the time is saying something back. But besides just responding, the agent can also perform actions, which are specified in a separate python file. For example, when the user asks for a restaurant nearby, the action *search_restaurant* is called, and it uses an API to extract restaurant details, to then give these to the user. An example of the start of a conversation is shown in Fig. 5.

The conversation altogether is a 'story'. Stories are pre-defined storylines the developer can create. These stories are the training data for Rasa Core. The training data for Rasa NLU are the intents. They are both trained using a neural network in Keras, based on an LSTM. This can be adjusted if necessary. The training data was created manually and is made available with the toolkit.

Cozmo4Resto & Rasa: Since the goal of Cozmo4Resto is to suggest restaurants nearby the user, the main action of Cozmo was to get the coordinates of the user (via their profile), find restaurants nearby of the type the user wants, and, if required, also give the address of the restaurant of choice. For the sake of simplicity, and for this application, we set the user location to a fix value and we propose only a single suggestion. Therefore we had two actions: *action_search_restaurant* and *action_give_address*. However, in a conversation people don't immediately ask what they want. The conversation usually has an introduction first (most often some sort of greeting). After

⁹https://github.com/Franck-Demoncourt/ASR_benchmark

¹⁰<https://rasa.com/docs/rasa/>

TABLE III
THE CHANGES MADE DURING TESTING.

round	change
1	added stories
2	if cuisine not recognised: ask another
3	added action_other_suggestion
4	if cuisine is 'anything': give random restaurant
5	adjusted fallback method
6	added entities

this, the robot can ask the user a question leading to the goal of the conversation. When the goal is reached, the conversation comes to an end, with some kind of goodbye-utterance. This results in six intents: greet, goodbye, wantdinner, cuisine, affirm, deny. Wantdinner let's the agent know that the user wants a suggestion to eat somewhere. If there is no cuisine suggested already, the agent will ask for the type of restaurant. This will then lead to a 'cuisine'-response from the user. After receiving the type of cuisine, it is put in a Slot, so the actions can 'grab' this when needed. Cozmo will then look for a restaurant, with the help of the Zomato API, an API to extract information of restaurants¹¹. Since we work with Python, the zomathon library was used¹². When a restaurant with the right cuisine was found, Cozmo gave it back to the user, by saying 'What do you think of Abc, a restaurant that serves xyz?'. The user could then either 'affirm' or 'deny' this restaurant. If denied, the agent should give another suggestion, however, this is not implemented yet. When affirmed, the agent asks if the user needs the address, and if they affirm, the agent gives it, using the action_give_address. After this, the goal is reached, and the conversation will close with a goodbye (and a 'bon appetit!' from Cozmo).

Evaluation: To test if the dialog management system was working properly, nine eINTERFACE participants were asked to chat with the system in a simple command line interface. They were instructed to get information on a restaurant in the neighbourhood but were not informed about how to do this, to make the user's utterances as free and unguided as possible. When they felt the conversation was complete, or were stuck and could not go any further, they informed the researcher. The average conversation was about 4,5 minutes. Afterwards, they filled out an evaluative questionnaire, with five questions on conversational fluency from Mirnig et al. [3], put on a 7-point Likert scale, and an open question about what problems occurred. We iteratively improved our dialog strategy based on the results of the questionnaires and the conversation which was recorded in the system's logs. Therefore, after the first three conversations and then after each one, the system was improved and tested again. An example of an improvement is an adjustment to the fallback method ("Sorry, I did not understand that."). This went on until the participants did not seem to run into any problems. The utterances of the users also became new training examples for the system. The changes made after each round are shown in table III.

Implementing it in the toolkit: To have the DM working within the toolkit, it needed to be able to run outside the

TABLE IV
MEAN OPINION SCORE OF THREE PARTICIPANTS OF SENTENCES SYNTHESIZED WITH DIFFERENT TTS SYSTEMS

MOS	IBM API	gTTS	SOTA batched	SOTA unbatched
P1	2.80 ± 0.10	3.24 ± 0.16	3.70 ± 0.16	3.88 ± 0.13
P2	3.43 ± 0.23	3.28 ± 0.25	3.10 ± 0.36	3.63 ± 0.26
P3	2.25 ± 0.40	2.30 ± 0.31	2.10 ± 0.37	2.90 ± 0.43
All	2.83 ± 0.19	2.94 ± 0.18	2.97 ± 0.24	3.47 ± 0.19

command line, and to be able to handle a JSON input file (from ASR), run the DM, and output a JSON file (to TTS). This was done by creating what is called a 'connector'-file. This contains the specifications on the input channel and a blueprint (from sanic¹³) on how to handle the input, namely how to send it to Rasa Core and retrieving Cozmo's responses. Since the toolkit was not entirely done by the end of eINTERFACE'19, the DM was made operable by connecting it to a Google Assistant.

F. Text-to-Speech (TTS)

As for ASR, some companies provide APIs for synthesizing speech from a text. Among them, *gTTS* is a python library allowing to use Google Translate built-in synthesizer. IBM provides the Watson TTS API¹⁴.

One of the best state-of-the-art (SOTA) Open Source implementations in terms of naturalness so far for TTS is the joint implementation of Tacotron [4] and WaveRNN [5] systems in PyTorch¹⁵. Tacotron generates a mel-spectrogram from text and WaveRNN generates the corresponding waveform sample by sample from the predicted mel-spectrogram. WaveRNN is able to produce a very natural sounding audio wave but generating the signal sample by sample with a recurrent relationship is still slow. This implementation proposes a way to accelerate generation of a sentence, called *batched*, by generating segments of the signal output of a sentence in parallel. The segments have to be concatenated together via a windowing process. This technique allows faster generation but leads to a chopped signal which is not the case of the *unbatched* generation.

For subjective evaluation, the 20 first sentences of harvard sentences¹⁶ were synthesized. Then three people evaluated them subjectively in terms of naturalness by assigning a score between 1 and 5. The Mean Opinion Score was then computed for each system.

Table IV shows the results of the MOS test for each participant and each system.

The synthesis duration is also an important aspect to consider since this module will be integrated in a HAI application where the agent has to respond in real-time. Concerning the SOTA systems, the durations of generation of mel-spectrograms and waveforms using a GPU GeForce GTX 1080 Ti are summarized in Table V. The order of magnitude

¹¹<https://developers.zomato.com/api>

¹²<https://github.com/abhishtagatya/zomathon>

¹³<https://sanic.readthedocs.io/en/latest/sanic/blueprints.html>

¹⁴<https://www.ibm.com/watson/services/text-to-speech/>

¹⁵<https://github.com/fatchord/WaveRNN>

¹⁶<http://www.cs.columbia.edu/hgs/audio/harvard.html>

TABLE V
DURATION (IN SECONDS) OF GENERATION OF MEL-SPECTROGRAM (ABBREVIATED MEL) WITH TACOTRON AND WAVEFORM (ABBREVIATED WAV) WITH Wavernn.

	mel	wav batched	wav unbatched
Mean	0.147419	12.771765	55.560221
Std	0.015291	0.173554	6.270252

for mel-spectrogram generation is 0.1 seconds. While for the waveform synthesis, it is one to several tens of seconds. In the unbatched mode, the standard deviation is much bigger because it depends more on the length of the sentence.

Therefore for a real-time interaction application one of the APIs is best suited since they both obtained reasonable and similar results on the MOS test and run relatively fast.

IV. COZMO4RESTO FRAMEWORK IMPLEMENTATION

For this application, the Behavior framework (client) will be connected with each module corresponding to the current state using the messaging library ZeroMQ¹⁷ (for other application, the toolkit allows the use of other messaging systems). The input and output of each module are in the JSON format containing three values: the data needed by the module as input or returned by it as output, the current state and the next state.

The state machine describing Cozmo's behavior in the case of the Cozmo4Resto application is detailed Fig. 6. But, due to the eNTERFACE workshop's time constraints, the platform was evaluated using a simpler application which is described as follows:

- state listening: A user's speech is converted into text using an ASR
- state thinking: keywords are mapped to other words in a dictionary playing the role of a very simplified dialog management system.
- state speaking: the words from the dialog management are sent to a TTS system to be synthesized state listening: the system goes back to the ASR

The aforementioned is provided to the reader for testing¹⁸.

V. CONCLUSION

In this paper, we report on the advancement in our project, presenting our goal for an open-source HAI toolkit and its application in our Cozmo4Resto project. In the future we will focus on integrating all the modules in a real-time application and test it in subjective experiments. Finally a generic library will be released as a first version of our open-source toolkit.

In future works, we will focus on implementing the behavior described in Fig. 6 by implementing the modules described previously into the platform. The whole system will then be tested in subjective evaluations by asking participants to first interact with Cozmo and then grading different aspects of the interaction like how well did the agent "understand" the

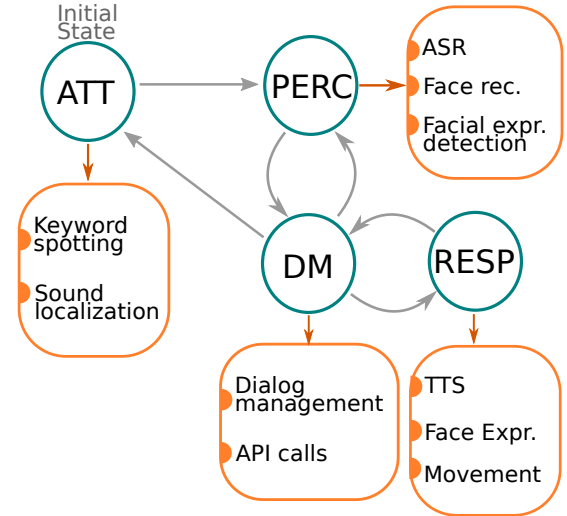


Fig. 6. State machine describing Cozmo's behavior for the Cozmo4Resto application. ATT: attention state which is linked to modules performing keyword spotting and sound localization-it is the initial state of the behavior, PERC: perception state which is linked to modules such as speech recognition (ASR) and face recognition, DM: the dialog management module which can also interact with API to harvest data from the web or control Cozmo directly, RESP: the response state generates a reaction to the user such as synthesized speech or a generated movement.

requests made by the user, the accuracy of the responses, the delay between the questions and reactions.

We will use the toolkit to create other HAI applications with platforms other than Cozmo serving us as agents such as 3D avatars [6].

REFERENCES

- [1] D. R. Wright, "Finite state machines," *Carolina State University*, p. 203, 2005.
- [2] F. Dérmoncourt, T. Bui, and W. Chang, "A framework for speech recognition benchmarking," in *Interspeech*, 2018, pp. 169–170.
- [3] N. Mirnig, A. Weiss, G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, B. Granström, and M. Tscheligi, "Face-to-face with a robot: What do we actually talk about?" *International Journal of Humanoid Robotics*, vol. 10, no. 01, p. 1350011, 2013.
- [4] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.
- [5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.
- [6] K. El Haddad, F. Zajega, and T. Dutoit, "An open-source avatar for real-time human-agent interaction applications," in *Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction*, 2019.

¹⁷<https://zeromq.org/>

¹⁸<https://github.com/kelhad00/hai-toolkit>

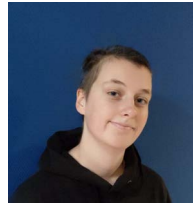


Kevin El Haddad is a teaching assistant and Ph.D. student at the University of Mons. His Ph.D. work currently focuses on the use of nonverbal and affective expressions in human-agent interactions. In 2016, he was a visiting researcher at the Trinity College of Dublin. In 2018, he spent 4 months as a lab associate at Disney Research (Los Angeles, California) working on AI and Human-agent Interaction systems. His research interests include machine learning, affective computing, human-agent interactions and signal processing. He lead 3 previous

eNTERFACE projects.



Noé Tits obtained his Master of Electrical Engineering specialized in Signals, Systems and Bio-engineering in June 2017. His Masters thesis was done in the University of the Basque Country in Bilbao (Spain) in the Aholab laboratory specialized in speech processing. His experience also count research projects in the field of electrical engineering such as simulations of heating of cables and electromagnetic fields in cable glands (Laborelec, GDF Suez), motion analysis (eNTERFACE workshop, Numediart Institute of UMONS), singing voice analysis and Medical Image Processing. In december 2017, No obtained a grant from the FNRS to pursue a doctorate at the Numediart Institute of UMONS. His current research focus on the application of Deep Learning techniques for controlling the emotional expressiveness in Text-to-Speech Synthesis.



Ella Velner is a PhD student at the University of Twente. She holds a MSc in Information Systems from the University of Amsterdam in 2019. Prior to that, she obtained a BA in Communication and Information Sciences from the VU Amsterdam. Her PhD focuses on responsible child-robot interaction, with a special interest in voice. Other interests are linguistics, machine learning, and human-media interaction.



Hugo Body is a Masters student in Artificial Intelligence and Smart Communications at the University of Mons (Graduating in 2021). He worked on an audio-based localization project for human-robot interaction applications for his bachelor degree in 2019.

Developing a Scenario-Based Video Game Generation Framework: Preliminary Results

Elif Surer ⁽¹⁾, Mustafa Erkayaoğlu ⁽²⁾, Zeynep Nur Öztürk ⁽³⁾, Furkan Yücel ⁽⁴⁾, Emin Alp Bıyık ⁽⁵⁾,
Burak Altan ⁽⁶⁾, Büşra Şenderin ⁽⁷⁾, Zeliha Oğuz ⁽⁸⁾, Servet Güner ⁽⁹⁾, H. Şebnem Düzgün ⁽¹⁰⁾

⁽¹⁾ Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

⁽²⁾ Department of Mining Engineering, Middle East Technical University, Ankara, Turkey

⁽³⁾ Department of Computer Engineering, Bilkent University, Ankara, Turkey

⁽⁴⁾ Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

⁽⁵⁾ Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

⁽⁶⁾ Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

⁽⁷⁾ Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

⁽⁸⁾ Department of Psychology, Bilkent University, Ankara, Turkey

⁽⁹⁾ Department of Mining Engineering, Middle East Technical University, Ankara, Turkey

⁽¹⁰⁾ Colorado School of Mines, Brown Hall 268, CO 80401, USA

elifs@metu.edu.tr, emustafa@metu.edu.tr, nur.ozturk@ug.bilkent.edu.tr,
furkanyucel.arch@gmail.com, emin.biyik@metu.edu.tr, burak.altan@metu.edu.tr,
busra.senderin@metu.edu.tr, zeliha.oguz@ug.bilkent.edu.tr, e119605@metu.edu.tr,
duzugun@mines.edu

Abstract—Emergency training and planning provide structured curricula, rule-based action items, and interdisciplinary collaborative entities to imitate and teach real-life tasks. This rule-based structure enables the curricula to be transferred into other systematic learning platforms such as serious games — games that have additional purposes rather than only entertainment. Serious games aim to educate, cure, and plan several real-life tasks and circumstances in an interactive, efficient, and user-friendly way. Although emergency training includes these highly structured and repetitive action responses, a general framework to map the training scenarios' actions, roles, and collaborative structures to game mechanics and game dialogues, is still not available. To address this issue, in this study, a scenario-based game generator, which maps domain-oriented tasks to game rules and game mechanics, was developed. Also, two serious games (i.e., Hospital game and BioGarden game) addressing the training mechanisms of Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNe) domain, were developed by both the game developers and the scenario-based game generator for comparative analysis. The results show that although the game generator uses higher CPU time, memory usage, and rendering time, it highly outperforms the game development pipeline performance of the developers. Thus, this study is an initial attempt of a game generator which bridges the CBRNe practitioners and game developers to transform real-life training scenarios into video games efficiently and quickly.

Index Terms—Serious Games, Game Scenarios, Video Game Generator, CBRNe, Emergency Training.

I. INTRODUCTION

SERIOUS gaming [1] [2], the umbrella term describing the video and board games having additional goals rather than only entertainment, is widely used in several domains such as health [3], defense [4] [5], and education [6] [7]. CBRNe is an acronym for Chemical, Biological, Radiological, Nuclear, and Explosives and recent research on this domain focuses on personnel training, emergency planning and organizing of field, tabletop, simulation, and serious gaming exercises for preparedness [8].

One of the use cases of serious gaming in emergency planning is on firefighter training. In a study by Heldal [9], firefighter training was examined by using serious games and tools. To do so, qualitative questionnaires and observations on two use cases (i.e., ship evacuation in Baltic Sea and railway accident with cyanide leakage) were used to analyze the impacts of serious gaming on non-users. Results showed that serious games would be useful in emergency training situations, and in-depth training scenarios and evaluation methods were necessary.

In another study, Lukosch et al. [10] performed the steps of the traditional design process with the contributions of the end-users. The primary purpose of the study was to check if the simulations could be used to train situational awareness skills, and the end-users' participation demonstrated the positive

impact of using simulations. However, the main limitation of the study was not having a game-scenario based approach, and future research would focus on this aspect while creating virtual agents.

The use of virtual reality simulation was also a common topic in the literature. Ingrassia et al. [11] focused on testing and comparing performances of 56 medical students during mass casualty triage in real-world and virtual reality (VR). The results showed that VR and live simulation were both useful in improving the accomplishments of the medical students. Ragazzoni et al. [12] also focused on VR training's medical aspect where the objective was to increase the staff safety in life-or-death risks. Hybrid simulation for infection control and Ebola treatment were also successfully performed virtually, and the results demonstrated that awareness of the health personnel increased.

Serious gaming in CBRNe has been a recent topic, and there are some misinterpretations on the definitions and core concepts, such as the misuse of the words 'game' or 'simulation'. To overcome these misinterpretations, a pre-development survey was developed [13] to be used before implementing the serious games of the European Network Of CBRN Training CEnters (eNOTICE) project [14]. In the pre-development survey, 24 questions were asked to the practitioners and experts of CBRNe under the following subgroups: 1) Participant's video gaming background, 2) Participant's knowledge on serious games, and 3) Participant's expectations on eNOTICE serious games. Results from 14 CBRNe professionals showed that the majority of the participants were highly positive on using serious games in CBRNe and provided open concepts, suggestions, and guidelines to develop serious games for CBRNe domain [15].

In this study, a scenario-based game generator, which maps linear real-life scenarios to serious games with training objectives, was developed. The main focus of this study is the CBRNe domain, where the roles, tasks, and goals of the participants are clearly defined. The main objectives of the games are as follows: 1) Providing a tool for additional training, 2) Synergy building, and 3) Transporting a different domain and a new concept to the CBRNe community. Two of the games that were developed by both the scenario-based game generator and by the game developers were based on the real exercises that were performed during eNOTICE project's joint activities in Nimes (France) and Brussels (Belgium). The comparative analyses regarding CPU usage, rendering, memory usage, and game development pipeline on the generator-based and developer-based games were also performed.

II. MATERIALS AND METHODS

In this section, details of the scenario-based game generator (Figures 1-3), two serious games that were developed by both the game designers and the game generator, and their evaluation are explained in detail.

1) *Scenario and Task Definitions*: The theme of the scenario, active players, location, and interaction mechanisms were collected in advance from the practitioners. Workflow and state diagrams were used to create a detailed scenario

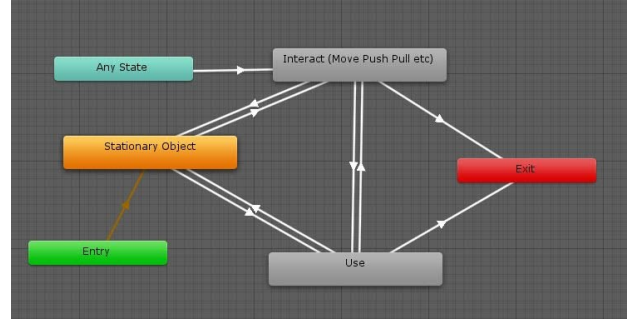


Fig. 1. Initial interaction mechanism including Entry and Exit States, a Stationary Object interacting (i.e., Move, Push, Pull, etc.) and using stationary objects.

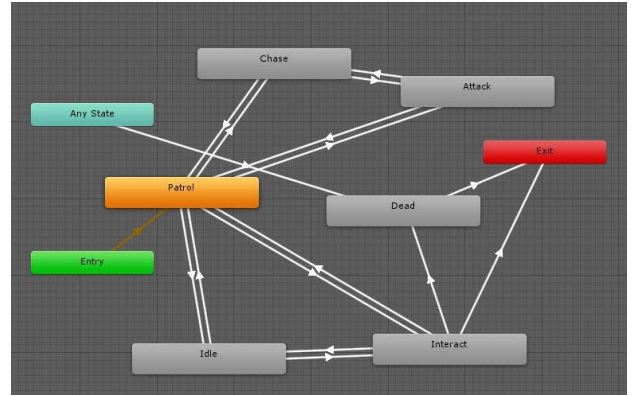


Fig. 2. In an attack scenario, chasing, attacking and interacting use cases are modeled.

where different roles and entities communicate with each other. In the beginning, two different CBRNe scenarios that were based on real practices were mapped to the workflow and state diagram structures. One of the scenarios is based on the subset of the BioGarden exercise (i.e., linear version of the scenario in which the players do not change the flow of the events), which was held in June 2018 in Belgium as part of eNOTICE joint activities. The other scenario is based on the Nimes exercise, which was held in France in January 2018 again as part of eNOTICE joint activities.

Defined scenarios and interaction mechanisms were mapped to game ideas, linear game stories, and interaction mechanisms. Interaction mechanisms were converted into concrete tasks and user roles. A generic system, built on top of the initial scenario definitions, was conceptualized and implemented. Then, the generated system was fine-tuned with goals, feedback measurements, and score adaptations.

The scenarios of the exercises were designed by different institutions such as fire departments, research centers, and hospitals. Thus, breaking the scenarios down into actions and events was a crucial step so that the game mechanics, reward mechanisms, and scoring could be systematized. Also, different roles in the scenarios were assigned to different player types so that the active role of the player and the role of non-player characters (NPCs) were defined.

Before starting to implement the Hospital game, a detailed survey, which was briefly mentioned above, was conducted

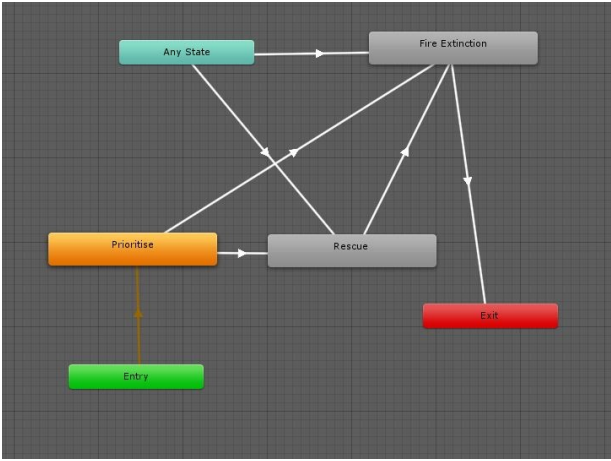


Fig. 3. Fire fighting scenario includes prioritizing the steps, extinguishing the fire and rescuing the affected people.

on 14 professionals in CBRNe field —7 of them being game players [15]. The scope and the purposes of the study were as follows: 1) Learning the user’s gamer profile, 2) Understanding the user’s perspective on serious games, 3) Retrieving the expectations of the user, 4) Clarifying the differences between the video/serious games and simulations, and 5) Asking for suggestions. The initial results of the survey were positive and encouraging. The participants’ gamer profiles involved playing strategy games and multiplayer games to learn new skills and to relieve stress. After the detailed analysis of the results, a tutorial mode was added to the initial game prototype.

2) *Scenario-Based Game Generator*: Scenario-based game generator is developed in combination with Unity software’s Animator Controller tool and is composed of four different components: 1) Main Code, 2) Control Code, 3) Transition Code, and 4) User Interface. Main Code is where the state definitions and structures —the definitions of the final consequences of the actions— are initialized. Control Code works as a mechanism to form and map action methods and their related states. Transition Code is the link where the game generator works with Unity’s Animator Controller. Finally, additional user interface mechanisms such as feedback, scores, and health points are added and grouped under the User Interface component (Figure 4).

While developing the scenario-based game generator tool, the following steps are executed: 1) Creating an environment, 2) Defining state diagrams, 3) Creating animations, 4) Adding basic artificial intelligence (AI) to states, 5) Resolving player and NPC interactions and 6) Adding basic AI for interactions.

In this scenario-based game generator, only linear scenarios, where the decision making of players do not modify the outcomes of the actions, are implemented. The scenario-based game generator is used to generate the duplicates of Hospital and BioGarden games. First, the game scenarios are tested using simple 3D game objects such as cubes, spheres, and capsules (Figures 5 and 6). Then, after checking that the scenario works correctly, initial 3D game objects are automatically replaced with game assets using tag information of the assets.

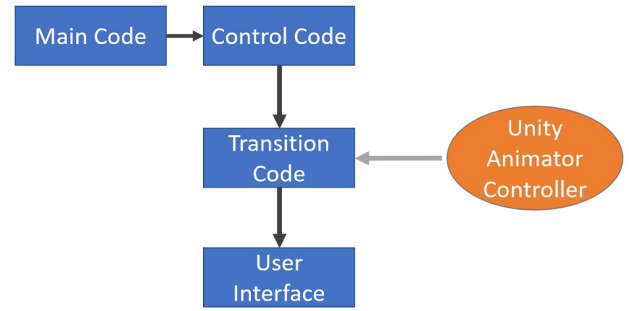


Fig. 4. Structure of the scenario-based game generator.

3) *Hospital Game*: Hospital Game is based on the Nimes scenario, which was performed during the eNOTICE joint activity in January 2018. The main purpose of this scenario is training medical staff for CBRNe circumstances. In this game, the player learns taking security measures such as using gloves, using masks, blocking the entrance of the hospital, and applying decontamination procedures. Players can play different roles, such as a doctor, nurse, and secretary. It is based on a linear scenario, and when the players make wrong choices, they lose game points (Figures 7 and 8).

This game was developed by a second-year Middle East Technical University (METU) Multimedia Informatics program student and the same game scenario was also given to the scenario-based game generator. The initial results of both environments were compared. In both versions of the games, Quadart’s Hospital Lowpoly pack [16], which provides several realistic modular assets, was used.

4) *BioGarden Game*: BioGarden game is based on the eNOTICE joint activity, which was played in Belgium in June 2018. Although it had a nonlinear scenario, only linear parts of the scenario were implemented so that a comparison with the scenario-based game generator would be possible. In the scenario, there were different laboratories with different structures and responsibilities. The role-playing part was composed of decontamination, role assignment, and evaluation (Figures 9 and 10).

As in the case of the Hospital game, BioGarden game was developed by a second-year METU Multimedia Informatics program student, and the same game scenario was also given to the scenario-based game generator, and the initial results were compared. In both versions of the games, 3LB Games’ Low Poly laboratory pack [17], which provides several realistic models, textures, and diffuse maps, was used.

III. RESULTS

In this section, performance outcomes of the scenario-based game generator and the two serious games that were developed by the game developers were compared in terms of CPU usage, rendering time, memory usage, and game development pipeline. All the tests were performed on a laptop having Intel Core i7 9750HQ CPU, 16GB RAM, and NVIDIA GeForce GTX 1660TI graphics.

Table I and Table II present the comparison on CPU usage (i.e., game generator’s output vs. developer-based game), Table

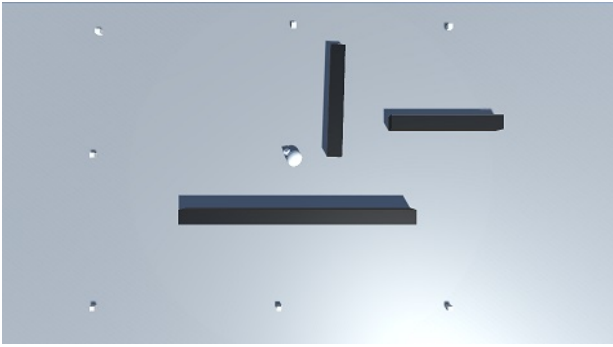


Fig. 5. Initial tests of the scenario-based game generator were performed on simple game objects.

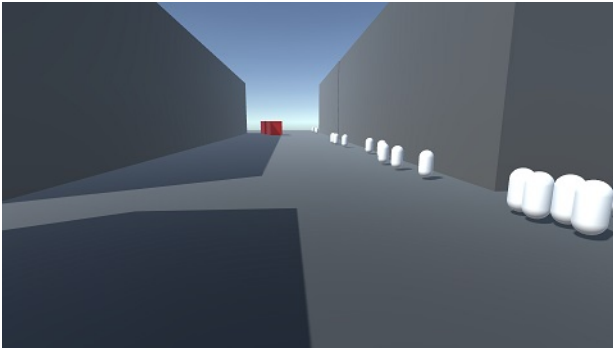


Fig. 6. Initial tests of the scenario-based game generator were performed on simple game objects such as cubes and capsules.



Fig. 7. Screenshots from the Hospital game and a demo of Dialogue menu.

III and Table IV on memory usage, and finally, Table V and Table VI present a comparison using rendering parameters. The rendering profile used the SetPass Calls, Draw Calls, Total Batches, Triangles and Vertices as parameters. SetPass parameter is defined as “the number of rendering passes” [18], a Draw Call as a “call to the graphics API to draw objects” [18] and Batch as a “package with data that will be sent to the GPU” [18] on the Unity’s Render Profiler page [18].

It took three weeks to develop and implement the scenario-based game generator. The most time-consuming part was the state transitions and handling the outcomes of the actions. After the game generator was built, it took three and a half hours to generate the Hospital game and four hours to generate



Fig. 8. Screenshots from the Hospital game and a demo of the interaction mechanism.



Fig. 9. The interior design of the Clandestine lab from the BioGarden game and menu interactions.



Fig. 10. The interior design of the Clandestine lab from the BioGarden game.

TABLE I
CPU USAGE OF THE HOSPITAL GAME GENERATED BY THE SCENARIO-BASED VIDEO GAME GENERATOR VS. HOSPITAL GAME

CPU Usage	Hospital (Generator)	Hospital Game
CPU	9 ms	4 ms

TABLE II
CPU USAGE OF THE BIOGARDEN GAME GENERATED BY THE SCENARIO-BASED VIDEO GAME GENERATOR VS. BIOGARDEN GAME

CPU Usage	BioGarden (Generator)	BioGarden Game
CPU	22.6 ms	8.5 ms

the BioGarden game using the game generator.

The development of the original BioGarden game took 25

TABLE III
MEMORY USAGE OF THE HOSPITAL GAME GENERATED BY THE
SCENARIO-BASED VIDEO GAME GENERATOR VS. HOSPITAL GAME

Memory	Hospital (Generator)	Hospital Game
Used Total	1.03 GB	0.45 GB
Reserved Total	1.32 GB	0.63 GB
System Memory Usage	1.96 GB	1.38 GB

TABLE IV
MEMORY USAGE OF THE BIOGARDEN GAME GENERATED BY THE
SCENARIO-BASED VIDEO GAME GENERATOR VS. BIOGARDEN GAME

Memory	BioGarden (Generator)	BioGarden Game
Used Total	0.62 GB	0.28 GB
Reserved Total	0.92 GB	0.49 GB
System Memory Usage	1.65 GB	1.27 GB

TABLE V
RENDERING RESULTS OF THE HOSPITAL GAME GENERATED BY THE
SCENARIO-BASED VIDEO GAME GENERATOR VS. HOSPITAL GAME

Rendering	Hospital (Generator)	Hospital Game
SetPass Calls	106	136
Draw Calls	252	298
Total Batches	217	243
Triangles	463.9K	504.9K
Vertices	319.5K	361.1K

TABLE VI
RENDERING RESULTS OF THE BIOGARDEN GAME GENERATED BY THE
SCENARIO-BASED VIDEO GAME GENERATOR VS. BIOGARDEN GAME

Rendering	BioGarden (Generator)	BioGarden Game
SetPass Calls	1826	2200
Draw Calls	2584	3007
Total Batches	2584	3007
Triangles	3.1M	3.2M
Vertices	2.3M	2.4M

days in total: one week for the scenario clarification and role assignment; one week for the text-based decision mechanisms, nine days to finish the user interface and menus and two days to add the assets to the game.

The original Hospital game was first refactored from the initial prototype, which took one week. Then, it took another two weeks adapting the new scenario, merging different game modes, and dialogue generation.

IV. DISCUSSION

In this study, real-life exercise scenarios of two eNOTICE joint activities were developed by game developers as well as a scenario-based game generator —developed during this study. All the frameworks used the same scenario-to-game mechanics mapping. All four versions of the games (i.e., developed by the game developer vs. generated by the game generator) were compared in terms of CPU usage, memory usage, rendering, and game development timeline perspectives. Scenario-based game generator's CPU usage, memory usage, and rendering time were higher when compared with the developer-based

games. The reason for the game generator's higher resource usage was the complex structure of the scenario-based game generator, tag search, and communication with the Unity's Animator Controller.

The rendering performance results of both versions of the games were very similar because the working principle of the game generator was not dependent on the visual contents of the games. Although the game generator used higher memory, CPU, and rendering time, its game development timeline efficiency highly outperformed the game developers (i.e., four hours vs. three weeks). This is a highly promising outcome that will enable further exercise scenarios to be mapped into games in a short period of time. This outcome can benefit the practitioners in two ways: 1) Visualizing the action-state diagrams of the exercise so that they can see the flaws or unassigned roles of their exercises, and 2) Having a rapid game prototype which becomes a fast, interactive testbed and training tool.

The proposed game generator framework will be extended using state machines so that nonlinear scenarios can also be generated quickly. Also, use case scenarios will be adapted to VR environments [19]. The players will interact with their surroundings in the VR environment, achieve their goals, interact with other users and receive feedback regarding the success of their outcomes, which will enable us to build a detailed training environment where training scenarios can easily be modified and played in two different settings: on computers and using VR headsets. All the game versions (i.e., developer-based and game generator-based) will be played by the users, and a comparative analysis will be performed. Finally, the performance, technology acceptance, immersion, and usability outcomes of the proposed system will be tested on participants and practitioners. Besides collecting game-related parameters such as interaction time and score, standard questionnaires on usability [20] and technology acceptance model [21] will also be applied to users. This tool will also be used to develop new prototype games for the future joint activities of the eNOTICE project till 2022.

V. CONCLUSION

In this study, a scenario-based video game generator, which targets the scenarios in CBRNe domain, was developed. This initial version of the game generator used linear scenarios that were based on the joint activities of the eNOTICE project. The effectiveness of the game generator was tested in comparison with two serious games, which were developed by the game developers. Even though the performance of the game generator lacked on the rendering, memory usage, and CPU usage aspects, it highly outperformed the game development pipeline of the game developers. This is a promising result that will enable the practitioners to visualize their scenarios while also generating prototype games rapidly so that the training of CBRNe personnel will be enriched. This current version of the game generator will be improved with usability tests, adaptation to VR and feedback of the CBRNe personnel so that an end-to-end and easy-to-use serious game generator for the CBRNe field will be provided.

ACKNOWLEDGMENTS

This study was developed during the 15th Summer Workshop on Multimodal Interfaces (eNTERFACE'19) which was held at Bilkent University, Ankara, Turkey between the dates of July 8 and August 2, 2019. This framework is fully supported by European Network Of CBRN TraIning Centers (eNOTICE) project funded under EU H2020 (Project ID: 740521). The authors would like to thank Dr. Olga Vybornova (Center for Applied Molecular Technologies, Université catholique de Louvain, Brussels, Belgium) and Prof. Gilles Dusserre (IMT Mines Alès, Alès, Languedoc Roussillon, France) for their help in scenarios of BioGarden and Hospital games, respectively. The authors would also like to thank Oğuzcan Ergün (Multimedia Informatics, Middle East Technical University, Ankara, Turkey) for his help and feedback on VR instrumentation.

REFERENCES

- [1] R. Santos, C. Magalhaes, L. F. Capretz, J. C. Neto, F. Q. B. da Silva, A. Saher, *Computer games are serious business and so is their quality: particularities of software testing in game development from the perspective of practitioners*, 2018. arXiv preprint arXiv:1812.05164 <https://arxiv.org/abs/1812.05164>.
- [2] D. R. Michael and S. L. Chen, *Serious games: Games that educate, train, and inform*, Muska & Lipman/Premier-Trade, 2005.
- [3] M. Pirovano, E. Surer, R. Mainetti, P. L. Lanzi, N. A. Borghese, *Exergaming and rehabilitation: A methodology for the design of effective and safe therapeutic exergames*, Entertainment Computing, 14, 55-65, 2016.
- [4] T. Susi, M. Johannesson, P. Backlund, *Serious games: An overview*, 2007.
- [5] D. Crookall, *Serious games, debriefing, and simulation/gaming as a discipline*, Simulation & gaming, 41(6), 898-920, 2010.
- [6] R. Tinati, M. Luczak-Roesch, W. Hall, *An investigation of player motivations in Eyewire, a gamified citizen science project*, Computers in Human Behavior, 73, 527-540, 2017.
- [7] V. Curtis, *Motivation to participate in an online citizen science game: A study of Foldit*, Science Communication, 37(6), 723-746, 2015.
- [8] XVR Simulation — Incident command training tool for safety and security, (September, 2019). Retrieved from <https://www.xvrsim.com/en-cn/>.
- [9] I. Høldal *Simulation and serious games in emergency management: experiences from two case studies*, 22nd International Conference on Virtual System & Multimedia (VSM), pp. 1-9. IEEE, 2016.
- [10] H. Lukosch, T. van Ruijven, and A. Verbraeck, *The participatory design of a simulation training game*, Proceedings of the Winter Simulation Conference, Winter Simulation Conference, 2012.
- [11] P. L. Ingrassia, L. Ragazzoni, L. Carenzo, F. L. Barra, D. Colombo, G. Gugliotta, F. Della Corte *Virtual reality and live scenario simulation: for training medical students in mass casualty incident triage*, Critical Care, 16(1), p.P479, 2012.
- [12] L. Ragazzoni, P.L. Ingrassia, L. Echeverri, F. Maccapani, L. Berryman, F.M. Burkle, F. Della Corte *Virtual reality simulation training for Ebola deployment*, Disaster medicine and public health preparedness, 9(5), pp.543-546, 2015.
- [13] User Expectations Questionnaire for eNotice Serious Game (2019, November). Retrieved from <https://ec.europa.eu/eusurvey/runner/eNoticeSeriousGamePreDevelopment>.
- [14] European Network Of CBRN TraIning CEnters – Official Website (2019, November). Retrieved from <https://www.h2020-enotice.eu>.
- [15] E. Surer, T. B. Atalay, D. Ç. Demirhan, H. Ş. Düzgün, *Serious Gaming in CBRNe Domain: A Survey on User Expectations, Concerns and Suggestions*, 3rd International Conference CBRNE - Research & Innovation, Nantes, France, p.43, 2019.
- [16] Unity Asset Store - Hospital LowPoly, (2019, September). Retrieved from <https://assetstore.unity.com/packages/3d/environments/hospital-lowpoly-82552>.
- [17] Unity Asset Store - Low Poly Laboratory Pack, (2019, September). Retrieved from <https://assetstore.unity.com/packages/3d/environments/low-poly-laboratory-pack-47677>.
- [18] Rendering Profiler - Unity Documentation (2019, November). Retrieved from <https://docs.unity3d.com/560/Documentation/Manual/ProfilerRendering.html>.
- [19] E. Surer, *Developing a Scenario-Based Video Game Generation Framework for Virtual Reality and Mixed Reality Environments*, he 15th Summer Workshop on Multimodal Interfaces (eNTERFACE'19) Project Proposal, pp.1-3, 2019. Retrieved from http://web3.bilkent.edu.tr/enterface19/wp-content/uploads/2019/03/p03_description.pdf.
- [20] J. Brooke, *SUS-A quick and dirty usability scale*, Usability evaluation in industry, 189(194), 4-7, 1996.
- [21] V. Venkatesh and F. Davis, *A theoretical extension of the technology acceptance model: Four longitudinal field studies*, Management science, 46(2), 186-204, 2000.



Elif Surer Elif Surer received her Ph.D in Bioengineering in 2011 from the University of Bologna. She received her M.Sc. and B.Sc. degrees in Computer Engineering from Boğaziçi University in 2007 and 2005, respectively. She is currently working as an Assistant Professor at the METU Graduate School of Informatics' Multimedia Informatics program. She is funded by the H2020 project eNOTICE as METU local coordinator as of September 2017 and also collaborates as a researcher in several interdisciplinary national and EU-funded projects. Her research interests are serious games, virtual/mixed reality, reinforcement learning and human and canine movement analysis.



Mustafa Erkayaoğlu Asst. Prof. Erkayaoğlu, holding a PhD degree in mining engineering and a minor degree in Systems and Industrial Engineering earned from the University of Arizona, is an expert in business intelligence in the mining industry with a broad academic and practical experience. He has a proven track of teaching experience on mining engineering courses and his expertise also extends to project management, continuous improvement, and performance measurement skills acquired in the projects he worked as a Business Intelligence (BI) expert.



Zeynep Nur Öztürk Zeynep Nur Öztürk is studying on her Bachelor's degree in Computer Science at Bilkent University. During her Bachelor's degree, she studied basics of artificial intelligence and programming games. Her focus is on game development with Unity, Unreal Engine and Java.



Furkan Yücel Furkan Yücel received his B.Arch. degree in Architecture from Bilkent University in 2019, where he studied generative algorithms and computational design in architecture. He is a first-year Master's student at METU Multimedia Informatics program.



Emin Alp Bıyık Emin Alp Bıyık received his B.Arch degree in Architecture from Middle East Technical University (METU) in 2018, where he studied creative coding and generative design in architecture. He is a first-year Master's student METU Multimedia Informatics program. His research interests are game development, virtual/mixed reality, generative systems and multimedia arts.



Zeliha Oğuz Zeliha Oğuz is a second-year psychology student at Bilkent University. She has been working at Bilkent University's National Magnetic Resonance Research Center (UMRAM) as a research assistant. The research is about understanding gene-environment interactions by using neuroimaging, genetic and psychological analysis and multimedia arts.



Burak Altan Burak Altan currently studies at METU Graduate School of Informatics' Multimedia Informatics program. He received his B.Sc. degree in Computer Engineering from Başkent University in 2017. He currently works at Ekin Teknoloji as a software engineer where his work includes front-end web development, Android development and database management. His research interests are serious games, virtual/mixed reality, and AAA game development.



Servet Gürer Servet Gürer received his BS.c degree in METU Mining Engineering Department and continues his Master's degree in the same department. After receiving his degree, he worked as a mining engineer in different companies and then became an occupational safety specialist. Since 2011, he has been working as a labor inspector at the Ministry of Family, Labor and Social Services. He is a member of TMMOB Chamber of Mining Engineers and Labor Inspectors Association. He served as the 44th board member of the Chamber of Mining Engineers.

He works in the fields of 3D modeling, animation, game development, virtual reality.



Büşra Şenderin Büşra Şenderin is a second-year Master's student at METU Multimedia Informatics program. Her current research focuses on serious games and virtual reality. She obtained her BS.c degree in Computer Engineering from Yıldırım Beyazıt University in 2018.



H. Şebnem Düzgün Dr. H. Şebnem Düzgün is Professor and Fred Banfield Distinguished Endowed Chair in Mining Engineering at Colorado School of Mines, Golden, USA. She also holds a joint appointment in the Department of Computer Science at Mines. Her recent research areas involve quantitative risk and resilience assessment for mining hazards and geohazards, big data analytics, Earth observation in geosciences, virtual/augmented/mixed reality (VR/AR/MR) and serious gaming for technical training and collaborative decision making.

Exploration of Interaction Techniques for Graph-based Modelling in Virtual Reality

Adrien Coppens ⁽¹⁾, Berat Bicer ⁽²⁾, Naz Yilmaz ⁽³⁾ and Serhat Aras ⁽²⁾

⁽¹⁾ Software Engineering Lab, University of Mons, Mons, Belgium

⁽²⁾ Department of Computer Engineering, Bilkent University, Ankara, Turkey

⁽³⁾ Cognitive Science Program, Bogazici University, Istanbul, Turkey

Corresponding author: adrien.coppens@umons.ac.be

Abstract—Editing and manipulating graph-based models within immersive environments is largely unexplored and certain design activities could benefit from using those technologies. For example, in the case study of architectural modelling, the 3D context of Virtual Reality naturally matches the intended output product, i.e. a 3D architectural geometry. Since both the state of the art and the state of the practice are lacking, we explore the field of VR-based interactive modelling, and provide insights as to how to implement proper interactions in that context, with broadly available devices. We consequently produce several open-source software prototypes for manipulating graph-based models in VR.

Index Terms—Human Computer Interaction, Virtual Reality, 3D User Interface, Graph Editing, Graph-Based Models, Interactive Modelling, Parametric Modelling.

I. INTRODUCTION

The overall goal of this eNTERFACE'19 project is to explore multi-modal interactions for manipulating graph-based models, i.e. visual models whose basic structure can be represented in the form of graphs, in Virtual Reality (VR). To approach the problem with different views, our team is composed of people with different backgrounds (computer science, computer engineering, architectural design and cognitive science).

We chose to work on architectural design as a case study, more specifically on parametric modelling. Recent work [1] has enabled “mesh streaming” from Grasshopper¹, a popular parametric modelling tool, to Virtual Reality, and identified benefits or visualising a geometry in such a context.

When designing with Grasshopper, an architect works in a visual programming language that relies on models based on directed acyclic graphs (DAGs), as an underlying representation. In such a graph, edges contain either standard parameters (e.g. numbers, booleans) or geometries that nodes process in order to output other geometries. The generated result can then flow into the graph and be used as input for other nodes. A complete architectural model can be designed that way, with the intended final geometry being typically output by a sink node. Figure 1 depicts such a DAG for a simple parametric model, whose purpose is to draw a cube defined by 2 corner points, P_1 at the origin and P_2 at (4, 4, 4) (since the same “side length” parameter is used for all 3 additions). The resulting

geometry (the output of the sink node i.e. “Box”) is shown on the side of that figure.

While [1] allows the user to modify parameter values within the VR environment, we here enable full editing of the graph-based model i.e. users can add, remove or move nodes, and add or remove edges from it. The resulting changes to the model can be saved back to the original Grasshopper-specific format, thanks to a framework that was developed prior to the workshop.

Although our work focuses on such architecture-oriented models, (most of) our findings will apply to other contexts where working within a VR environment makes sense and where graph-based modelling is required. The interaction techniques we rely on are not even limited to DAG-based models, and examples of application domains include software engineering (e.g. several types of UML diagrams), transportation network design, robotics (e.g. path planning) and the entertainment industry (e.g. scene graphs for games, animation design). These fields are, at least potentially, dealing with 3D content and could therefore benefit from an immersive design environment, that matches the content’s dimensionality (e.g. visualising a 3D animation in VR while designing it from the same immersive environment, seems beneficial).

II. RELATED WORK

A. Context of the case study: architectural modelling

Architectural design tools evolved over the course of the discipline’s history. While paper drawings and scale models are still relevant today, they are now accompanied by 2D and 3D modelling software, with limited support for Augmented Reality (AR) and Virtual Reality (VR). Regardless of the exact tools in use, design activities tend to follow a well-defined process that goes from task analysis to the final product. Howard et al. described such a process [2] by combining literature reviews on both engineering design and cognitive psychology. They describe a creative design process composed of four steps, as depicted in Figure 2. After analysing the needs for the task in hand, a designer works on creating (often multiple) conceptual designs. Based on a concept, the designer then builds up a structure (i.e. shapes the abstract concept into a concrete design) in the embodiment phase. As indicated by the arrows in that figure, it is quite common to go back and forth between those three steps. Once satisfied with

¹<https://www.grasshopper3d.com>

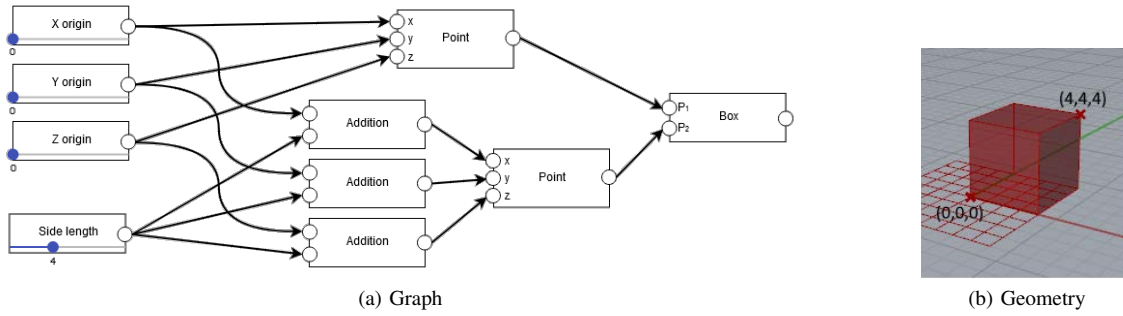


Fig. 1. DAG (a) for a simple parametric model defining a cube, with the corresponding output geometry (b).

the result, the designer polishes the geometry and produces communication documents for physically building the model.

Most tools that attempt to bring AR/VR technologies in use for architectural modelling focus on the visualisation part and typically provide limited (e.g. texture switching) live editing features (or none at all) that would allow the VR user to modify the geometry whilst immersed into the virtual environment. Examples of such tools include IrisVR Prospect² and Twinmotion³. There have been several efforts to provide annotation and sketching capabilities in a VR context, some of which were targeted at architectural design activities (e.g. [3] that relies on a tablet to control a 3D cursor used for sketching), but those are “only” conceptual design activities.

More recently, the MARUI plugin⁴ has enabled users to modify models directly within a VR environment, therefore tackling the problem of bringing embodiment/detailed design activities into VR. While this clearly appears to be a big step towards the technology’s integration for modelling activities, it does not currently offer parametric modelling capabilities.

In order to fill that void, the work we previously mentioned [1] describes a proof-of-concept application that enables parameter sharing in addition to geometry streaming. This means that an architect can interact with the parametric model’s parameter values and see what the resulting effect is on the geometry, all from within the VR environment. Although this is a good starting point, there is a clear need to expand the tool’s capabilities with full model (graph) editing.

B. 3D programming languages

Numerous 3D programming languages for virtual 3D environments have been designed, most of which rely on three-dimensional dataflow diagrams to define programs. Examples from the early 90s include CUBE/CUBE-II [4][5], a functional language, and Lingua Graphica [6], who translates from/to C++ code.

While the previous examples were indeed designed with virtual environments in mind, they were never adapted to immersive displays (e.g. VR). On the other hand, Steed and Slater implemented an immersive system [4] that allows users to define object behaviours whilst immersed, once again through dataflow graphs. The system could be used to design

animations or interactive applications, that conveniently also took place within the virtual environment.

More recently but with similar goals in mind, Lee et al. developed an Augmented Reality (AR) system [5] to define the behaviour of scene objects for AR applications. Once again, being able to develop and test a target application concurrently was pointed out as a clear benefit.

Since architectural geometries are three-dimensional, we believe designing them from within a VR system would be beneficial to the architect. Even though we therefore are in a 3D context, parametric models (the graphs) are usually two-dimensional, at least that is how they are laid out in the most popular desktop-based applications. While we also want to explore solutions to capitalise on the third dimension offered by VR environments, we will do so in a limited fashion (e.g. use a node’s height to indicate how close it is to the sink) rather than turning the parametric models into “unconstrained” 3D graphs. Indeed, our VR-based system relies on a table metaphor (i.e. the model to be manipulated is placed on a virtual table) to maintain some consistency with desktop tools. We decided to use this metaphor in order to follow the Information Visualisation community’s advice on not making use of unmotivated 3D layouts [6].

C. Interaction techniques for 3D environments

1) *Direct manipulation*: Interacting efficiently within a 3D immersive environment is likely to require input methods that differ from the traditional mouse and keyboard combination. Manipulation can be subdivided into 4 tasks: selection, positioning, rotation and scaling [7]. While a complete modelling environment needs to afford proper interaction mechanisms for all of these tasks, we will focus here on selection and positioning, the primary needs for our case study.

One way of categorising immersive interaction techniques is based on isomorphism: isomorphic approaches will preserve a natural one-to-one mapping between input actions and their resulting effect, whereas non-isomorphic techniques afford non-realistic interactions and can even be based on “magical” or “virtual” tools. Quite often, these techniques either rely on a *touching* metaphor a *pointing* metaphor.

As for touch-based techniques, the user must reach the target object’s position to interact with it. An isomorphic example of such a technique would be to track a physical controller with six degrees of freedom and map its position

²<https://irisvr.com/prospect/>

³<https://www.unrealengine.com/en-US/twinmotion>

⁴<https://www.marui-plugin.com>

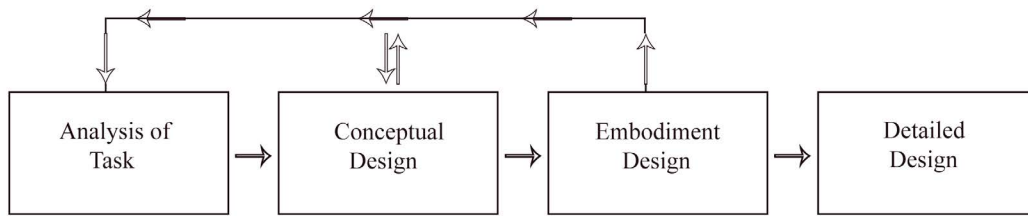


Fig. 2. The creative design process as described by Howard et al. [2].

and orientation to a virtual cursor, so that interaction with an object can be achieved by pressing a button when the virtual cursor collides with the target object. When the tracking zone is limited in space compared to the “interactable” area, non-isomorphic mappings are typically used to mitigate these limitations. Examples include the Go-Go [8] and PRISM [9] techniques, that both rely on non-linear mappings between the cursor and the tracked objects’ motions.

The pointing metaphor allows to mitigate space-related limitations, since aiming at an object is sufficient to start interacting with it. Similarly, the user can reposition that object by pointing towards a target position. Typically, a tracked controller’s position and orientation defines a “laser beam” that is used to select and manipulate objects. Techniques based on that metaphor differ in how the position and orientation of the controller affect the laser beam, and which object(s) are selected based on it. The simplest version is often called ray-casting, where the controller defines a line segment and its intersection with the environment defines the target object or position. More complex techniques allow users to bend the line (e.g. using Bzier curves [10]) to mitigate limitations related to occlusion, or make use of selection volumes instead of simple rays, which can result in easier selection and can potentially allow multiple objects to be selected. The selection volume size can be static (e.g. the spotlight technique [11] that relies on a selection cone) or dynamic (e.g. the aperture technique [12] that expands on spotlight by allowing users to control the spread angle of the cone).

2) *Speech Recognition*: In addition to the aforementioned interaction techniques, specific actions can be greatly simplified by relying on speech recognition. Since specifying an arbitrary position or selecting an existing object is easily done with a pointing metaphor, speech recognisers are often used in a multimodal context when applied to 3D selections or manipulations (e.g. the “Put-That-There” metaphor [13]).

A plethora of speech recognition tools and techniques are available for use. Some can be used offline whereas others are based on an online service; they can either listen to the user in a continuous manner or await specific actions (e.g. a button press or an API call). An important distinction between speech engines is whether (and how much) they restrict potential input. Free speech recognisers can output any text whereas directed dialogue [14] systems are limited to a set of predefined words or commands. Directed approaches can mostly be found in two forms: keyword-spotting solution that extract specific words; and grammar-based tools that produce phrases defined by specific rules.

Our use case would benefit from vocal commands such as

“Add component X” or “Add slider with value 7”, to create new (potentially valued) nodes in the graph. Even though free speech and keyword-based approaches could be used for that purpose, they would not guarantee that a valid output is returned by the speech recogniser and would require manual parsing of that output (which sequences of words or keywords are valid, and what action they correspond to). Grammar-based engines therefore seem to be the best option as only valid vocal commands, with regards to the grammar, can be recognised. The challenge therefore moves from post-processing the result to correctly defining the rules of the grammar. The remaining of this section will consequently present the Speech Recognition Grammar Specification (SRGS). This is a W3C standard⁵ that describes a grammar format. Similarly to the grammars from compiler theory, a SRGS grammar describes a set of rules composed of tokens, using either an XML or a BNF-based (Backus-Naur Format) syntax.

SRGS grammars can be augmented with Semantic Interpretation for Speech Recognition (SISR⁶) tags that contain ECMAScript (JavaScript) code to be executed when the corresponding grammar rule is matched. Those tags are typically used to assign values to a matched rule (e.g. a boolean value can be set to `true` when the matched text is “yes”, “ok” or “yeah”), especially when handling numbers (e.g. saying “three” assigns the value 3 to the variable `outValue`; and saying “thousands” multiplies `outValue` by 1000).

III. TARGETED ACTIONS AND INTERACTION POSSIBILITIES

The two basic elements of a parametric design graph are nodes and edges, with each node containing any number of input and/or output ports. In order to interact with the graph and in addition to the ability to modify parameter values, a user should be able to add, move and remove nodes. User should also be capable of adding and removing edges as well as moving and scaling the viewpoint. By viewpoint, we mean the part of the graph that is currently visible to the designer.

Considering the devices we have at our disposal for this workshop (HTC Vive and its controllers⁷, Leap Motion⁸ and Kinect⁹), Figure I helps in visualising the exploration space i.e. which interaction techniques are, were or can be used to realise those actions. That table implies that we classify these interaction techniques based on the modality they rely on and whether they interact with the target object directly.

⁵<https://www.w3.org/TR/speech-grammar/#S1>

⁶<https://www.w3.org/TR/semantic-interpretation/>

⁷<https://www.vive.com/us/product/vive-virtual-reality-system/>

⁸<https://www.leapmotion.com/>

⁹<https://developer.microsoft.com/en-us/windows/kinect>

TABLE I
TARGETED ACTIONS AND OPTIONS FOR INTERACTION TECHNIQUES.

Techniques Actions		Modality			Interaction type			
		6-DoF controller	Hands	Speech	Direct	Indirect	Isomorphic	Non-isomorphic
Node	Add	P_1	P_2	P_1	P_1, P_2	P_1, P_2	P_1, P_2	P_2
	Remove	P_1	P_2		P_1, P_2		P_1, P_2	P_2
	Move	P_1	P_2		P_1, P_2		P_1, P_2	P_2
Edge	Add	P_1	P_2		P_1, P_2		P_1, P_2	P_2
	Remove	P_1	P_2		P_1, P_2		P_1, P_2	P_2

A. First prototype: Vive controllers

Our first prototype (P_1 on Table I) relies on the default controllers provided with the HTC Vive. They are tracked with 6 degrees of freedom (DoF), meaning that their 3D position and rotation are both tracked simultaneously. For this prototype, we chose to explore an isomorphic interaction technique based on the grasping metaphor: the user simply touches the element he wants to have an interaction on, and presses a button to trigger the corresponding action. Figure 3 shows a user that is about to grasp a node in P_1 .

If that element is an edge, we simply remove it from the graph. If it is a node, we attach it to the controller (that type of interaction is often referred to as the grasping metaphor). The user can then either release it somewhere else on the graph (realising the “move node” action) or throw it away (“remove node” action). In order to add an edge to the graph, the user simply has to select two ports. After selecting the first port and prior to selecting the second one, a temporary line between the selected port and the controller is rendered so as to give feedback to the user on the port that has been interacted with. Note that adding an edge is prevented if that edge would create a cycle in the graph (since Grasshopper can only work with acyclic graphs).

A video demonstrating this prototype is available online¹⁰ and our codebase is hosted as open-source software on a GitHub repository¹¹.

B. Second prototype: Leap Motion

The goal of the second prototype (P_2 on Table I) is to mimic the isomorphic techniques from P_1 and explore non-isomorphic and indirect interaction with the Leap motion finger-tracking device. The sensor is composed of different cameras that give it the ability to scan the space above its surface, up to 60 cm away from the device.

In order to manipulate and control the basic elements of the graph-based model, we rely on gestures. Captured gestures determine what type of action to apply to these elements. We first started by recognising and categorising gestures captured from the Leap Motion, in real time. Events we recognise include “hover”, “contact” and “grasp” but we also can rely on a non-isomorphic gesture: “pointing” (using raycasting). Each

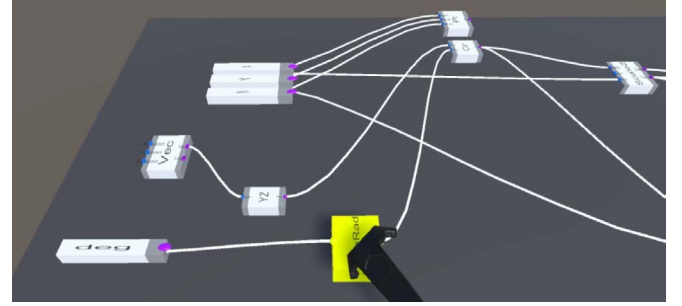


Fig. 3. A user grasping a node with our first prototype.

of these events/gestures can be detected separately for both left and right hands, simultaneously.

Unfortunately, P_2 's codebase and assets were not compatible with the framework developed for the workshop, and we consequently could not integrate the explored techniques into a fully working prototype.

C. Speech recognition

Since our prototypes are made with the Unity game engine, we can benefit from its built-in support for the Windows Speech Recognition API, that includes an XML-based SRGS grammar recogniser. We therefore defined a grammar that handles node creation. A user can indeed add a new node in the graph by saying “Add Component type”, where type is the node type. The list of valid node types for the grammar is dynamically generated at runtime, based on the component templates previously learned by the system (e.g. upon encountering a new component type when loading a model from a file).

Since we sometimes need to assign a value to a newly added component (e.g. “Add slider with value 7”), we also need to recognise numbers. Alphanumeric input in grammars is nontrivial [15] but we relied on an existing set of rules provided in the Microsoft Speech Platform SDK¹².

Lastly, as users may want to assign an arbitrary text value to a node (e.g. a panel component), we had to make use of a free speech recogniser, that starts listening to user input only when specific grammar rules have been processed. In the meantime, the grammar recognition engine is paused, and it only resumes when the user stops providing free speech input.

¹⁰<http://informatique.umons.ac.be/staff/Coppens.Adrien/?video=eNTERFACE2019>

¹¹<https://github.com/qdrien/eNTERFACE-graph-architecture>

¹²<https://www.microsoft.com/en-us/download/details.aspx?id=27226>

IV. EVALUATION

A. Setup

To evaluate our prototypes, we will develop different scenarios that focus on fundamental features of the system (e.g. add a certain type of node or remove a specific edge). After an introduction to VR to limit biases related to (un)familiarity with the technology, we will ask participants to go through equivalent scenarios with our VR prototypes and question them during the session to obtain live feedback (i.e. interview/demo approach [16]).

We will also ask participants to evaluate the usability of the system after each session, with a standard post-hoc questionnaire [16] SUS (System Usability Scale [17]). Combining these complimentary methods will allow us to benefit both from a deep level of live feedback and an overall evaluation of the system.

B. Evaluation Criterion

Different indicators of usability will be evaluated:

a) *Completion time for a specific task*: A different input method may perform better on a particular subset of actions, and we therefore need to make sure that proper interaction techniques are available.

b) *Error rate*: Quick completion only makes sense when the result matches the intended effect, so we will keep track of errors made during the experiments (e.g. the wrong edge has been deleted).

c) *Intuitiveness*: Thanks to the aforementioned SUS questionnaire, we will be able to measure how easy it is for users to guess what actions will produce the intended result. Should difficulties be identified, we will take corrective actions so as to reduce the “Gulf of Execution” [7].

V. FUTURE WORK

Since P_2 could not be integrated, we will have to pursue our efforts with regards to other interaction techniques before evaluating the prototypes as described in section IV, with architects and architectural students. We will also adapt our work to other use cases in Software Engineering (e.g. UML diagram editing) in the near future, so as to validate the genericity of our approach.

VI. CONCLUSION

After an introduction to the chosen use case’s context, this report described our work on parametric modelling in VR, with a focus on the Human-Computer Interaction aspects and a multimodal approach. Even though we focused on a specific use case (parametric architectural modelling), we believe our experiments and findings are beneficial for all graph-based modelling activities in VR. More contributions to the domain are needed to fully take advantage of VR as a visualisation technology.

REFERENCES

- [1] A. Coppens, T. Mens, and M.-A. Gallas, “Parametric modelling within immersive environments: Building a bridge between existing tools and virtual reality headsets,” *36th eCAADe conference*, 2018.
- [2] T. J. Howard, S. J. Culley, and E. Dekoninck, “Describing the creative design process by the integration of engineering design and cognitive psychology literature,” *Design studies*, vol. 29, no. 2, pp. 160–180, 2008.
- [3] T. Dorta, G. Kinayoglu, and M. Hoffmann, “Hyve-3d and the 3d cursor: Architectural co-design with freedom in virtual reality,” *International Journal of Architectural Computing*, vol. 14, no. 2, pp. 87–102, 2016.
- [4] A. Steed and M. Slater, “A dataflow representation for defining behaviours within virtual environments,” in *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*. IEEE, 1996, pp. 163–167.
- [5] G. A. Lee, C. Nelles, M. Billinghamurst, M. Billinghamurst, and G. J. Kim, “Immersive authoring of tangible augmented reality applications,” in *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2004, pp. 172–181.
- [6] N. Elmqvist. (2017) 3d visualization for nonspatial data: Guidelines and challenges. [Online]. Available: <https://sites.umi.acs.umd.edu/elm/2017/10/03/3d-visualization-for-nonspatial-data-guidelines-and-challenges/>
- [7] J. J. LaViola, *3d User Interfaces: Theory and Practice*, 2nd ed. Hoboken, NJ: Pearson Education, Inc, 2017.
- [8] I. Poupyrev, M. Billinghamurst, S. Weghorst, and T. Ichikawa, “The go-go interaction technique: non-linear mapping for direct manipulation in vr,” in *ACM Symposium on User Interface Software and Technology*. Citeseer, 1996, pp. 79–80.
- [9] S. Frees and G. D. Kessler, “Precise and rapid interaction through scaled manipulation in immersive virtual environments,” in *IEEE Proceedings. VR 2005. Virtual Reality, 2005*. IEEE, 2005, pp. 99–106.
- [10] A. O. S. Feiner, “The flexible pointer: An interaction technique for selection in augmented and virtual reality,” in *Proc. UIST’03*, 2003, pp. 81–82.
- [11] J. Liang and M. Green, “Jdcad: A highly interactive 3d modeling system,” *Computers & graphics*, vol. 18, no. 4, pp. 499–506, 1994.
- [12] A. Forsberg, K. Herndon, and R. Zeleznik, “Aperture based selection for immersive virtual environments,” in *ACM Symposium on User Interface Software and Technology*. Citeseer, 1996, pp. 95–96.
- [13] R. A. Bolt, “Put-That-There”: *Voice and Gesture at the Graphics Interface*. ACM, 1980, vol. 14, no. 3.
- [14] R. Pieraccini and J. Huerta, “Where do we go from here? research and commercial spoken dialog systems,” in *6th SIGdial Workshop on Discourse and Dialogue*, 2005.
- [15] Y.-Y. Wang and Y.-C. Ju, “Creating speech recognition grammars from regular expressions for alphanumeric concepts,” in *Eighth International Conference on Spoken Language Processing*, 2004.
- [16] D. A. Bowman, J. L. Gabbard, and D. Hix, “A survey of usability evaluation in virtual environments: classification and comparison of methods,” *Presence: Teleoperators & Virtual Environments*, vol. 11, no. 4, pp. 404–424, 2002.
- [17] J. Brooke *et al.*, “Sus-a quick and dirty usability scale,” *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.



Adrien Coppens is a PhD student in the Software Engineering Lab at the University of Mons, in Belgium, where he also received a MSc degree in Computer Science, in 2017. His thesis is co-supervised by the Faculty of Architecture and Urban Planning since its topic is about bringing Augmented (AR) and Virtual Reality (VR) technologies in use in the context of Computer-Aided Architectural Design (CAAD), with a particular focus on parametric modelling. As relying on those technologies requires new kinds of interfaces and interaction paradigms,

he is heavily interested in Human-Computer Interaction, especially within 3D immersive environments.



Naz Yilmaz Naz Buse Yilmaz received her BSc. Degree in Interior Architecture from Istanbul Technical University (ITU), in 2018. She started her graduate studies at Cognitive Science Program, Bogazici University and currently MSc. Cognitive Science student at Graduate School of Informatics, Middle East Technical University. Her research interests include perception, visual and spatial cognition and human computer interaction.



Berat Bicer Berat Bicer was born in Tekirdag, Turkey in 1997. He received his B. Sc. degree from Department of Computer Engineering of Bilkent University, Ankara in June 2019 and is currently pursuing his Master's Degree in Computer Engineering at Bilkent University, Ankara under supervision of Hamdi Dibeklioglu. He's currently working on multimodal deceit and online spam detection. His research interests include computer vision, human behavior analysis, and pattern recognition.



Serhat Aras Serhat Aras was born in Ankara, Turkey in 1996. He received his Bachelors Degree in Computer Science from Bilkent University, Ankara, in 2019. Currently, he is pursuing a career based on Software Development, Machine Learning and Computer Vision. He is also a researcher at Bilkent University under the supervision of Hamdi Dibeklioglu. His research interests include Pattern Recognition, Computer Vision, Human Behavior Analysis, Gesture Recognition and Edge Computing.

Spatiotemporal and Multimodal Analysis of Personality Traits

Burak Mandıra ^(1,*), Dersu Giritlioğlu ^(1,*), Selim Fırat Yılmaz ^(2,*), Can Ufuk Ertenli ^(3,*),
Berhan Faruk Akgür ⁽⁴⁾, Merve Kınıklıoğlu ⁽⁴⁾, Aslı Gül Kurt ⁽⁴⁾, Merve Nur Doğanlı ⁽⁵⁾, Emre Mutlu ⁽⁶⁾,
Şeref Can Gürel ^(7,8), and Hamdi Dibeklioglu ⁽¹⁾

- (1) Department of Computer Engineering, Bilkent University, Ankara, Turkey
(2) Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
(3) Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
(4) Department of Neuroscience, Bilkent University, Ankara, Turkey
(5) Department of Psychology, Yildirim Beyazit University, Ankara, Turkey
(6) Psychiatry Clinic, Etimesgut Şehit Sait Ertürk State Hospital, Ankara, Turkey
(7) Department of Psychiatry, Hacettepe University, Ankara, Turkey
(8) Department of Cognitive Neuroscience, Maastricht University, Maastricht, Netherlands

burak.mandira@bilkent.edu.tr, dersu@bilkent.edu.tr, selim.yilmaz@bilkent.edu.tr,
ufuk.ertenli@metu.edu.tr, faruk.akgur@bilkent.edu.tr, m.kiniklioglu@bilkent.edu.tr,
gul.kurt@bilkent.edu.tr, mervenurdgnl@gmail.com, mutluemre12@gmail.com,
scgurel@hacettepe.edu.tr, dibeklioglu@cs.bilkent.edu.tr

Abstract—Analysis of personality traits is a common study of interest for various fields including psychology, psychiatry, and neuroscience. In the recent years, with the improvements in computing power, it has also become a popular area of study in computer science. Recent machine learning and computer vision models are able to interpret behavioral cues, such as, facial expressions, gaze, posture, gesture, voice, and speech to analyze observable personality traits. Yet, accessible assessment tools are still substandard for practical use, not to mention the need for fast, accurate and reliable methods for such analyses to evaluate personality traits. In this study, we present spatio-temporal and multimodal approaches to estimate the Big Five personality traits from audio-visual cues and transcribed speech. Furthermore, developing a robust desktop application, we automate the data acquisition, and collect a new audio-visual database, namely Bilkent Personality Database, for the task of personality analysis. In contrast to available databases in the literature, Bilkent Personality Database includes video recordings of induced behavior in addition to speech videos. As well as systematically assessing the reliability of different behavioral modalities on Bilkent Personality and ChaLearn LAP First Impressions databases, we evaluate the discriminative power of induced behavior for personality analysis. Our experimental results show that the induced behavior indeed includes signs of personality.

Index Terms—Big-Five, Personality traits, Personality trait analysis, Machine Learning, Computer vision, Multimodal fusion, Facial expression dynamics, Gaze, Gesture, Pose, Speech

I. INTRODUCTION

PERSONALITY traits have been widely studied by psychologists and psychiatrists. Various theories and methods have been proposed to determine personality traits of individuals. Among many other personality analysis theories,

TABLE I
ASSOCIATED ADJECTIVES FOR THE FIVE PERSONALITY TRAITS

Factor	Adjectives
Agreeableness (AGR)	Appreciative, Forgiving, Generous, Kind, Sympathetic
Conscientiousness (CON)	Efficient, Organized, Planful Reliable, Responsible, Thorough
Extraversion (EXT)	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative
Neuroticism (NEU)	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying
Openness to Experience (OPE)	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests

trait-based ones are widely accepted [1]. According to Goldberg [2], the model consists of five independent traits, namely, openness to experience (people who are curious to experience new things and imaginative), conscientiousness (people who are dutiful and self-disciplined), extraversion (people who are gregarious and active), agreeableness (people who are tolerant and trusting), neuroticism (people who tend to notice threatening factors in non-threatening situations and are emotionally unstable). McCrae and John provide empirical and theoretical foundations of the Five Factor model: It integrates various personality constructs; it is comprehensive (provides a way of systematic exploration of the relations between personality and other phenomena) and it is efficient (providing a global description of personality with as few as five scores) [1]. Table I demonstrates some example adjectives for each of these traits.

*These authors contributed equally.

These traits are mainly associated with cognition, affect, and behavior, such as Conscientiousness being dominated more by behaviour, Neuroticism by negative affect alongside these kind of behaviours, Extraversion by both affective and behavioral, and lastly Openness and Agreeableness by cognitions [3]. [3] suggests that certain personality traits are more visible to eye than some. In this sense, traits as extraversion, conscientiousness or neuroticism would be more apparent as we observe an individual at first sight. Effect of the Big Five traits on emotions have been studied [4], yet, there is limited to none research that investigates possible models for finding implications of these traits. Therefore, in this study, optimal solution is to design a computational model that can accurately identify implications of certain traits and use such certain marks of these traits in further annotations to validate each other.

Earlier studies have repeatedly demonstrated that the personality traits affect clinical features, prognosis and treatment response of certain mental disorders such as depression [5], personality disorders [6], post traumatic stress [7], substance use, and addiction [8] as well as psychotic spectrum disorders [9]. Even though evaluation of personality traits holds high potential to be effectively used in clinical settings for the management of certain disorders, this is hampered by certain aspects of current evaluation methods like requirement of specific training for application and interpretation, employment of extra personnel, and high time expenditure [10]. Therefore, automated reliable computerized methods for personality trait assessment could potentially overcome such limitations and increase their utilization, enabling better management of mental disorders.

In this study, as well as using an existing database, we construct/collect a new one. Furthermore, we develop methods employing deep architectures to analyze audio-visual cues in the videos also with the help of the transcribed speech. The contributions of this study can be listed as follows:

- A new audio-visual database for personality analysis is collected, including 60 subjects.
- We present spatiotemporal models for the estimation of personality traits from multimodal cues.
- We systematically evaluate the reliability of several behavioral modalities for personality analysis.
- We analyze the relation between self-reported and observed (by experts) personality traits.
- Our results suggest that the induced behavior includes signs/cues of personality.

II. RELATED WORK

For the analysis of personality traits, there are mainly four modalities that researchers focus on: image-based, text-based, combination of image and text, and audio-visual modalities. Some researchers also combine audio, image and transcription of speech. In this section, we will overview such recent studies.

Cucurull *et al.* [11] build a combined image-and-text based personality trait model, where they use a dataset collected from Instagram. They investigate whether there exists a correlation between the images that users post based on accompanying

text and the Big Five personality traits. For data collection, they chose 22 words for each personality trait, and acquire images according to those words. About 1,100 images are selected for each word, which results in 121,000 images for training the model. They simplify the problem as a binary classification problem such that for each trait they classify subjects as being low or high, where they use multinomial logistic loss. They also propose all-in-one network that combines these five classifiers together and evaluate it for different loss functions. They employ Alexnet and ResNet-50 as feature extractors that are pre-trained on ImageNet. Their best model is ResNet all-in-one that achieves a classification accuracy of 71.9%.

In contrast to [11], Segalin *et al.* [12] use images that users liked. The problem is defined as a binary classification task, where authors distinguish between high and low classes by defining thresholds with quartiles. They also employ ImageNet pretrained AlexNet and VGG16 models, and fine-tune their parameters according to PsychoFlickr corpus. Their model achieves about 53% self-assessed and 60-70% attributed accuracy depending on the target trait. Their best model (VGG16) achieves an average accuracy of 55% on self-assessed traits and 68% on attributed traits.

Wei *et al.* [13] has won the first round of 2016 Looking at People (LAP) ECCV Challenge on the First Impressions track, which aims to recognize apparent personality traits from 15-second videos based on the Big Five personality traits. Along with the scores of personality traits, dataset also includes job-interview assessments that represents the possibility of getting an invitation to a job interview. Authors use the first version of the First Impressions dataset [14], that consists of 10,000 clips extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing a camera and speaking in English, to train and test their models. Authors propose Deep Bimodal Regression framework. There are separate models for image and audio features where they merge them via late fusion. They propose DAN+, an extension to Descriptor Aggregation Networks, that utilizes max and average pooling at two different layers of the CNN and concatenating the normalized values before feeding them to dense layers. They also fine tune the pretrained VGG-Face and ResNet networks. When it comes to their architecture for modeling voice, they employ log filter bank (logfbank) features and a single fully-connected layer with sigmoid activations. They fuse all models' features at the end to create an ensemble model and achieve 0.9130 mean score (one minus mean absolute error (MAE) loss) over five traits. However, the main disadvantage of their model is that it do not fully exploit the temporal information since about 100 images are extracted from each video by a sampling rate of 6 frames per second.

While Wei *et al.* [13] has won 2016 LAP Challenge on the first round at ECCV, Gurpinar *et al.* [15] has won the second round of the same challenge at ICPR. The method proposed in [15], utilizes audio, video, and scene features. It achieves 0.913 mean score ($1 - \text{MAE}$) for the five traits.

[16] presents the state of the art on First Impressions v2 dataset [14]. This model has won the ChaLearn 2017 Looking at People (LAP) CVPR/IJCNN Competition with being the only submission that surpasses the baselines. The

baseline models of the challenge include language, sensory (audio and image) and their combined modalities. Highest average scores achieved (in terms of $1 - \text{MAE}$) for sensory, language-based, and combined modalities are 0.9109, 0.8867, and 0.9118 (for the five traits), respectively. These models use ResNet-18 for sensory modalities with the late fusion which is the concatenation of their latent features after the global average pooling layers of the models. The baseline method for the language modality uses an embedding that represents transcripts as 4800-dimensional mean skip-thought vectors. [16] represents language (i.e., transcripts), audio, and visual data as described in [15], where they use various features from action units, deep facial features, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) to deep scene and acoustic features. The main difference between [15] and [16] is that Escalante *et al.* [16] first combine the modalities at feature level, and stacking the predictions of sub-systems to an ensemble of decision trees, obtaining an average score of 0.9172 for the five traits.

III. METHODOLOGY

To model and estimate the level of (observed) personality traits, we employ several methods on different modalities including facial appearance, action units, head pose & gaze, body pose, voice, and transcribed speech. Observed scores for each trait (normalized to $[0, 1]$ range) are used as labels. Details of modeling each of these modalities will be described in the following sections.

A. Facial Appearance

1) *Face Normalization*: As the first step of analyzing facial appearance, we detect/track 68 landmarks on facial boundary (17 points), eyes & eyebrows (22 points), nose (9 points), and mouth (20 points) regions in the videos using a state-of-the-art tracker, namely OpenFace [17] (see 1). Once the facial landmarks are obtained, facial image in each frame of the videos are normalized in terms of translation, rotation and scale to obtain frontal view of the faces.

The tracked 2D coordinates of the landmarks are first normalized by removing the global rigid transformations such as translation, rotation and scale. To shape-normalize facial texture, each face image is warped using piecewise linear warping so as to transform the X and Y coordinates of the detected landmarks onto those of normalized landmarks. Obtained images are then scaled and cropped around the facial boundary and eyebrows as shown in Fig. 2. As a result, each normalized face image has a resolution of 224×224 pixels. Notice that the deformations in the facial surface can better be interpreted since the normalized faces are directly comparable in a pixel-to-pixel manner.

2) *Modeling*: Once the normalized facial videos are obtained, we model the spatio-temporal patterns using two different deep architectures, namely by the 3D ResNext-101 [18] and by a Convolutional Neural Network, Gated Recurrent Unit combination (CNN-GRU).

Since our input is a facial video, our aim is to capture both facial appearance and facial dynamics. To this end,

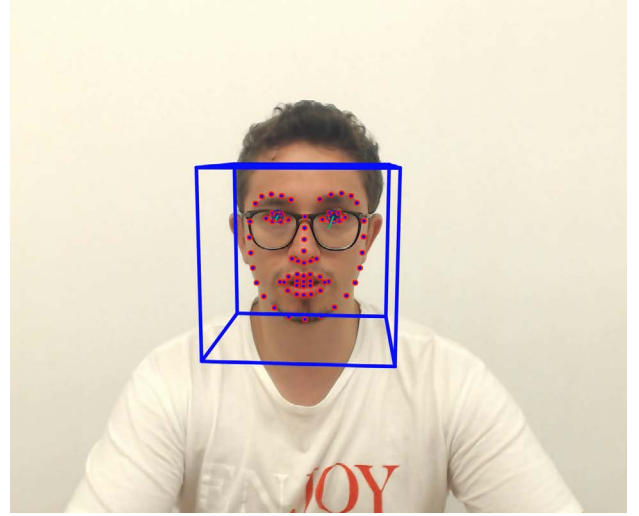


Fig. 1. Visualization of the facial landmarks, gaze direction, and head pose obtained from OpenFace.



Fig. 2. (a) A sample frame, and (b) its face-normalized version.

we opt for employing a CNN-based architecture that also takes into account the dynamics between the frames through spatio-temporal kernels. Therefore, we first use **3D ResNext** model to utilize temporality thoroughly. The novelty of the ResNeXt architecture [18] is the introduction of *cardinality* concept, which is a different dimension from deeper and wider. ResNeXt block introduces group convolutions (whose numbers are called cardinality), which divide the feature maps into small groups different than the original ResNet [19] bottleneck block. Xie *et al.* [18] shows that increasing the cardinality of 2D architectures is more effective than using wider or deeper architectures.

To model normalized facial videos, we fine-tune 3D ResNext-101 [20], which is pretrained on Kinetics dataset [21], starting from the third block (our preliminary experiments have shown that fine-tuning the third block is better than fine-tuning the fourth). We use random temporal sampling of 45 frames (RTS-45), which corresponds to 1.5 seconds, during training, and non-overlapping sliding window of the same size during test and validation. Window size is chosen among the values [30, 45, 60] through validation error. Finally, the last fully connected layer of the network is replaced with a linear regression layer and L1 loss is utilized.

CNN-GRU is employed as a second spatio-temporal deep

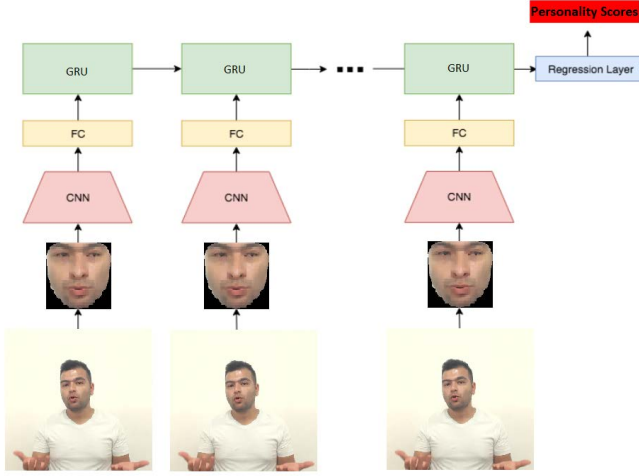


Fig. 3. CNN-GRU architecture, followed by a regression layer.

architecture for modeling facial videos. It is widely used [22], [23] in the literature, as it can model the spatial relations via CNN and temporal relations via the recurrent network at the same time. In our implementation, as shown in Fig. 3, AlexNet is used as the CNN module by connecting its FC7 to the GRU structure. In this way, 4096D spatial representation of facial images is fed to the temporal model. As the final layer linear regression with L1 loss is employed. The obtained model is trained in an end-to-end manner. We initiate the training from the pretrained weights of the original AlexNet, in order to accelerate the process and start from an effective set of parameters.

During training, average mean absolute error of the five traits is minimized for both 3D ResNext-101 and CNN-GRU models.

B. Facial Action Units and Head Pose & Gaze

1) *Feature Extraction*: To obtain measures for facial shape, displayed facial action units (AU), head pose, and gaze, we process the videos using OpenFace [17] as visualized in Fig. 1. In order to describe facial action units, we use the 18 AU occurrence and 17 AU intensity features provided by OpenFace. While the binary occurrence features indicate the presence of AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, and AU45, the intensity features (in the range of [0,5]) represent the intensity of the aforementioned AUs except AU28 (lip sucking).

To represent head pose, 3 degrees of out-of-plane rigid head rotations (i.e., pitch, yaw, and roll) in radians, and the 3D location of the head with respect to camera in millimeters are used. Finally, for describing the gaze, we employ the 3D gaze directions for both left and right eyes (yielding six feature values) together with the 2D coordinates of 28 eye landmarks for each eye (yielding 112 feature values). Then head pose and gaze features are concatenated, to be used as the representation of the head pose & gaze.

Obtained frame-level feature vector of each of these modalities can be used as a time step in temporal models. In other

words, each video can be represented by the multivariate time series of the aforementioned modality-specific feature vectors.

2) *Modeling*: The action unit and head pose & gaze features are modeled using two different models such as Long- and Short-term Time-series Network (LSTNet) [24] and Recurrent Convolutional Neural Networks (RCNN) [25], which combines the benefits of Long Short-Term Memory (LSTM) with CNN. Average mean absolute error of the five traits is minimized to train the models.

LSTNet model used in this study is a modified version of the original architecture [24]. We opt for LSTNet since the literature indicates that it achieves significantly better performance than various other time series models [24]. LSTNet extracts short term patterns and local dependencies via convolution through temporal dimension. The output of convolution layer is fed to the recurrent layer and the recurrent-skip layer. In recurrent and recurrent-skip layers, Gated Recurrent Unit (GRU) is used. Normally, GRU fails to capture very long-term dependencies due to gradient vanishing. Recurrent-skip layer captures long term and periodical information by processing the sequence with N skips, where a recurrent layer processes consecutive inputs with 1 skip-length. Then the output of recurrent and recurrent-skip layers concatenated and fed into the linear layer. We set the skip-length parameter to the number of frames per second. We apply dropout with a rate of 0.2 after convolution, recurrent, and recurrent-skip layers. The hidden dimension of convolution and recurrent layers are chosen as 100. In contrast to [24], we do not use the autoregressive component of LSTNet. We also do not use the tanh activation function at output since our target problem is regression. We train the network through optimizing the L1 loss via Adam optimizer with a learning rate of 0.001.

RCNN has been proposed in [25] for text classification. It uses Bidirectional Long Short Term Memory (BiLSTM) networks followed by max-pooling through temporal dimension. The output of the max-pooling layer is fed to the linear layer. Our RCNN's recurrent module consists of two BiLSTM layers. We set the dimension of all hidden layers of both backward and forward LSTMs as 256. Hidden dimension of linear output layer is set to 64.

C. Body Pose

We process the input videos using OpenPose [26] to track 25 landmark points on the joints (e.g., wrist and elbow), neck, and face as shown in Fig. 4. Obtained 2D coordinates of these landmarks are used as posture features to represent the general pose and structure of subjects' body, e.g. how they sit and move while answering questions and watching videos. Notice that apart from other visual features, this is the only modality where we focus not on the face, but the body/posture of the participant. 50-dimensional body features are then modeled with LSTNet as described in Section III-B2 so as to minimize the average mean absolute error of the five traits.

D. Voice

To represent the characteristics of voice, we compute a 34-dimensional feature vector from audio data of videos, including Mel Frequency Cepstral Coefficients, Chroma vector,

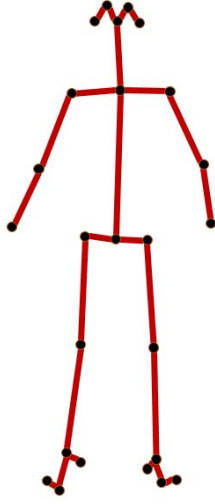


Fig. 4. The tracked body landmarks by OpenPose.

energy and entropy related features using pyAudioAnalysis framework [27]. Voice features are extracted for each 50 milliseconds of videos. Consequently, these features form the multivariate time series for describing the voice. Details of the feature extraction process can be found in [27]. Obtained features are modeled by LSTNet architecture (as described in Section III-B2) by minimizing the average mean absolute error of the five traits.

E. Transcribed Speech

Recent studies, e.g., [15] and [16], show that the use of language as an additional modality, improves the performance of estimating personality traits. In order to model language-based cues for personality analysis, we first transcribe the subjects' speech in videos using Google's Speech to Text API [28]. Since there may be more than one language spoken in the database, language is automatically detected. To make our model generalizable, an embedding obtained by a large corpora is used. We employ pytorch-transformers' implementation [29] of pretrained multilingual BERT model [30] to generate embeddings for each token in the transcripts. Notice that a multilingual model is essential since there may be more than one language in the database.

BERT model is applied to the whole transcript of each video, separately. BERT model infers the embeddings of each word in the transcript considering its context. Flow of our transcribed speech model can be seen in Fig. 5. BERT embeddings are used as input features to our model. For our multitask regression task, we use LSTNet (for details please see Section III-B2). Similar to the modeling of aforementioned modalities, average mean absolute error of the five traits is minimized during training.

IV. DATABASES

In our experiments, we employ two databases, namely Bilkent Personality Database, a new personality database that has been collected during this study, and the ChaLearn LAP

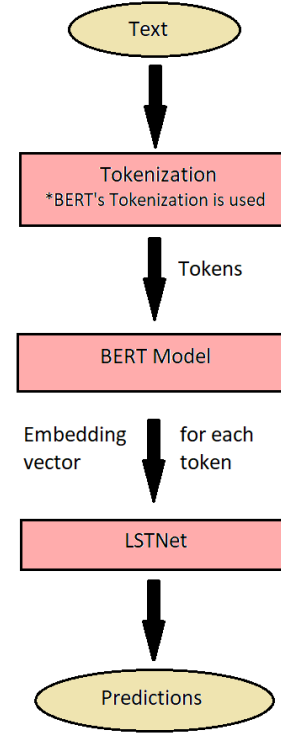


Fig. 5. Flow of modeling the transcribed speech.

First Impressions Database [14]. Below, these databases will be described in detail.

A. Bilkent Personality Database

One of the goals of this study is to investigate whether the personality traits can be estimated from induced audio-visual behavioral characteristics. To this end, we have video-recorded participants while they watch a set of videos, where each video has been chosen to be associated with one of the Big Five personality traits. In addition, we have recorded their answers to three questions. Bilkent Personality Database (BilPeD) includes recordings of 60 participants (37 females, 23 males) while they answer three questions, and watching 15 video clips: three video clips for inducing behavioral cues of each of openness, conscientiousness, extraversion, agreeableness, and neuroticism. The database includes self-assessed and observed scores for each the five traits.

To minimize the differences between sessions so as to obtain similar experience for different participants, we have developed and used a computer software rather than employing an interviewer during the data collection. However, before beginning the data collection, we have informed each participant about the experiment and the use of our software. After that the whole experiment and the data acquisition have been conducted automatically. Following sections will provide further details of BilPeD.

1) *Data Acquisition:* The software allows participants to choose their preferred language, either English or Turkish. Once a participant choose his/her preferred language, the software show three videos. In each video, a psychologist

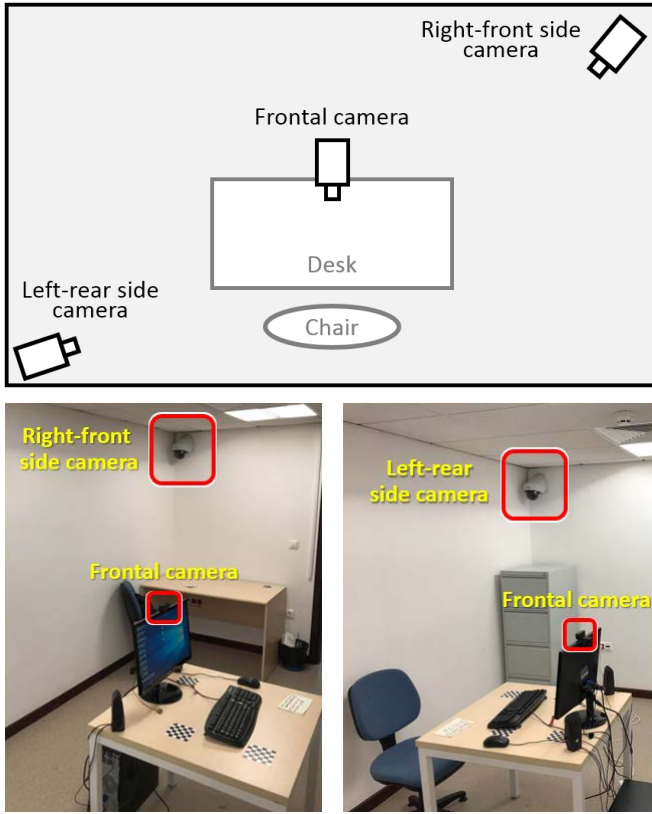


Fig. 6. Recording setup. Positioning of the right-front side, left-rear side, and frontal cameras.

ask a question (in the preferred language). The first question asks demographic information of the participant. The second question is about an experience of the participant while having an activity last time which he/she likes. In this way, the participant could specifically talk about a memory without thinking much on a certain one. The last question is about a time that the participant had solved a problem with his/her close other and how they managed it. After watching each question video, the participant is given 60 seconds to answer the corresponding question, with a 15 seconds countdown at the end in order to remind the remaining time. Then the video for the following question starts playing.

Once the participant completes answering the questions, he/she moves on to the second part of the experiment. In this part, to induce behavioral cues of personality traits, the software shows a set of video clips for approximately 15 minutes in total, including three videos for each of the five traits (15 videos in total). Duration of the videos varies approximately 30 to 60 seconds to obtain a proper response from the participant. Please notice that a large set of video clips have been chosen (by a consensus of three psychologists) in order to induce each of the five personality traits, and the software randomly chooses three videos from the corresponding set of videos for each participant.

The software records participants via three cameras. A Logitech C920 webcam is used as a frontal camera to record the facial expressions (which is attached to the monitor). Two wall mount cameras record the participants from right-

front and left-rear sides (with respect to the participant) to obtain pose and gesture information. The recording setup and sample frames captured by these three cameras can be seen in Fig. 6 and Fig. 7, respectively. While frontal videos have been captured with a resolution of 1920×1080 pixels at 30 frames per second, side-view videos have a resolution of 1280×720 pixels at 25 frames per second. Audio has been recorded with a sampling rate of 44100 Hz.

In the final stage of the experiment, the participants are asked to complete two ten-item questionnaires on seven-point scale, namely, "Ten Item Personality Inventory" (TIPI) and a questionnaire on close relationships ("Experiences in Close Relationships Inventory-Revised").

All the aforementioned steps of the experiment is handled by our software through an easy to use graphical user interface where the only focus is the monitor during the experiment, free of any other distractions.

2) *Choosing Video Clips to Elicit Behavioral Cues*: Picking the correct video that would tap into a specific personality trait is crucial, which might provoke a behavioral mechanism we may catch on. To this end, we investigate the traits and their prominent properties. Many descriptions of Big Five personality traits were used in defining which type of video clips we could pick.

For openness, we have chosen to focus on the curiosity and intellect aspect of this trait, since they would be easier to express. Recent studies suggest that openness is related to being a multicultural person, who tends to oversee the racial and ethnic differences of other people [31]. Therefore, we have decided to pick videos that includes activities, where participants presumably would not encounter on daily basis. Such videos would clearly display different cultures or religions (e.g. extreme sports, clips on different religions praying and their rituals).

For conscientiousness, we have picked videos that would show people's differences on academic performances and being diligent in each and every manner [32] (e.g. two students preparing for their exam: hardworking versus lazy).

For extraversion, we have deduced that individuals who have less fun interacting with others, would also be the ones that are introverts. Since individuals who have higher levels of extraversion tend to be more expressive, we assume that getting a reaction/response based on approval would be easier [33]. So, we have used clips that involve high physical stimulation in other people's presence (e.g. dance parties, performing to a crowd).

For agreeableness, we have chosen to focus on the social aspect on this trait. Therefore, we have decided to pick videos that would emphasize harmony with others [34]. Thus, we have used videos that include interpersonal encounters that is either harmonious with others or not (e.g. apologizing, disagreeing to anything without any logical basis).

For neuroticism, we have employed video clips that would look and make individuals feel like something bad is going to happen, inducing the disturbed thought processes of participants without actually making them feel that way. Therefore, if the corresponding participant is high on neuroticism continuum, he/she would expect something bad would happen more



(a)



(b)



(c)

Fig. 7. Sample frames obtained from (a) the right-front side, (b) left-rear side, and (c) frontal cameras.

than others (e.g. a house burning, glass falling down from a table without a reason). In this sense, we expect people to behave in a certain way when viewing these videos. Yet, [33] suggests that people that are high on neuroticism tend to be less expressive on their affect. So, it might be difficult to catch those expressions if the corresponding participant has high scores on neuroticism. As explained before, an introvert might be overwhelmed by over stimulation factors contained in the videos, whereas an extrovert might express positive affect with showing behaviors of blending in (e.g. bopping head to an upbeat song). We especially picked some scenes from movies and real life events for neuroticism that might trigger one's negative affect to further provoke disturbed thoughts.

3) *Annotation:* Obtained video recordings of participants were evaluated by three different psychologists in terms of personality traits. The psychologists discussed each personality trait of each participant until they achieved a 100% consensus on the final score. Participants' postures and facial expressions while they were watching the aforementioned video clips (e.g. their reactions/responses) were accounted for annotating the personality traits. Personality trait levels of the participants were annotated based on the relation/correlation between their responses and the target trait of the corresponding video. Seven-point scale was used for the annotation of the scores (1: very low; 7: very high).

For Openness trait, smiling and engagement with the video (more saccadic eye-movement without negative viewing) were pursued in individuals who are high in Openness. On the contrary, disgust-like facial responses were treated as low Openness. For Conscientiousness, high scores were given if the participants become disturbed after individuals being messy with their environments. If the participant shows engagement with the opposite type of behavior, he/she was rated low on Conscientiousness. For Extraversion, we again looked for engagement with the extrovert behaviors in the video clips. Other than this, we also looked for tapping of foot or swinging with the music when giving high scores. As opposed to these reactions, participants who gave disgust-like reactions or show discomfort were given low rating on Extraversion. For Agreeableness, we have expected individuals with high Agreeableness to remain calm while are shown an individual who is low on this trait. Others, who were showing discomfort, were rated low on this trait. Finally, for Neuroticism, individuals who were watching the video clips with significant amount of discomfort (e.g. squinting eyes, leaning back, etc.) even though scenes did not show any discomforting image, were rated high on Neuroticism.

If a participant did not show any signs of being in any given side of the spectrum of reactions while watching the (inducing) video clips, score for him/her was rated as 4 (neutral). If a participant showed any leaning to one side of the spectrum slightly, score for him/her was rated as 3 or 5 accordingly. If the participant showed considerable reaction (e.g. clearly displaying a certain response to the video), we rated his/her score as 2 or 6. Any extreme case of these spectra of behaviors was rated as 1 or 7, accordingly.

4) *Data Partitions:* As described above, BilPeD has recordings for durations of speaking (question answering) and for durations of watching video clips to induce cues of personality traits. These partitions of our database will be referred to as BilPeD-Interview and BilPeD-Induction, respectively, in the remainder of the paper. The interview partition includes 180 sessions (60 participants \times 3 questions), and the induction partition has 180 sessions for the induction of each trait, yielding 540 sessions in total (60 participants \times 3 inducing videos for each trait \times 5 traits). Notice that there are three synchronized videos, namely one frontal, and two side views, for each session.

B. ChaLearn LAP First Impressions Database

ChaLearn LAP First Impressions Database (FID) [14] contains 10,000 videos, splitted to training (6,000 videos), validation (2,000 videos) and test (2,000 videos) subsets. The subjects in the videos are looking at the camera and speaking in English, with varying environmental conditions. These videos are extracted from 3,000 different YouTube videos and labeled via Mechanical Turk, where the annotators rate the personality scores of each subject in terms of Big Five Traits, according to their movements, gestures, voice and appearance.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

In our experiments, two databases are used, i.e., Bilkent Personality Database (BilPeD) and the ChaLearn LAP First Impressions Database (FID). For the experiments on BilPeD, a 10-fold cross-validation scheme is used with randomly selected folds, ensuring that each subject appears only in one fold. In this way, we guarantee that there is no subject overlap between the training, validation and test sets. Random selection of the subjects to form the folds is processed once, and the obtained folds are used in all experiments for the sake of comparisons. In our experiments on FID, we the predefined training, validation and test sets of FID are used. Hyperparameters of the models are optimized on the validation set. Test results are reported in terms of mean absolute error (MAE). Notice that seven-point scale scores for each trait is normalized to the range of $[0, 1]$ in our experiments. Therefore, presented MAEs are also in the range of $[0, 1]$.

B. Observed versus Self-assessed Personality Scores

One of our goals in this study is to analyze and reveal the relations between the observed (expert-annotated) and self-assessed personality scores. To this end, we compute the distributions (histograms) of the observed and self-assessed scores for 60 participants in BilPeD, as shown in Fig. 8. Mean and variance of scores for each trait are also computed and reported in Table II. As the obtained results suggest, subjects tend to rate the traits, which are more socially desirable (e.g. traits that are perceived positive), higher than the observed scores. On the other hand, participants assess traits, which are less socially desirable, lower than observers do. This social desirability effect can especially be seen in the openness and the neuroticism. Participants overrate their openness by 39% and underrate their neuroticism by 22% compared to the observed scores.

In psychology research on personality traits, self-reported scores of an individual (e.g., TIPI scores) generally used more often than the observations, since they are fast and easy to collect and apply. However, people's ideas on their own personality and the manifestations of them may not always be coherent. As the results suggest, people tend to express themselves in a more socially desirable fashion in self-reports. Therefore, they may see themselves in a better way. For instance, although some individuals may score themselves high on openness, they may display contradicting behavioral

TABLE II
MEAN AND VARIANCE OF SELF-ASSESSED AND OBSERVED SCORES FOR THE FIVE PERSONALITY TRAITS OF 60 PARTICIPANTS IN BILPEd.

	Type	AGR	CON	EXT	NEU	OPE
Mean	Self-assessed	0.643	0.671	0.615	0.479	0.722
	Observed	0.594	0.614	0.567	0.608	0.517
Variance	Self-assessed	1.667	1.837	2.392	2.139	1.206
	Observed	1.012	0.816	1.440	1.494	1.523

TABLE III
MEAN ABSOLUTE ERRORS FOR THE USE FACIAL APPEARANCE.
* DENOTES FINETUNING ON BILPEd-INTERVIEW WITH AN INITIALIZATION USING WEIGHTS THAT ARE LEARNED ON FID.

Database	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	3D-ResNext	0.085	0.089	0.088	0.091	0.085	0.088
	CNN-GRU	0.097	0.105	0.101	0.102	0.101	0.101
BilPeD-Interview	3D-ResNext	0.169	0.129	0.217	0.181	0.196	0.179
	3D-ResNext*	0.153	0.109	0.186	0.163	0.176	0.158
	CNN-GRU	0.175	0.173	0.205	0.204	0.201	0.192
	CNN-GRU*	0.138	0.123	0.169	0.162	0.189	0.156

characteristics (e.g. squinting eyes, showing disgust-like facial expressions) in response to the videos that are linked with openness. Similarly, one may do the opposite for less socially desirable traits (e.g. neuroticism). Therefore, these biases should be taken into consideration in self-report scales. Based on these findings, the observed scores will used as data labels (for computational models) in the remainder of our experiments.

C. Assessment of Different Modalities

In this set of experiments, we evaluate the reliability of different modalities such as the facial appearance, facial action units, head pose & gaze, body pose, voice, and transcribed speech on BilPeD-Interview and FID for assessing the levels of personality traits. While results on FID are obtained using models that are trained on FID (training set), two set of test results are provided for BilPeD-Interview: (i) training on BilPeD-Interview, (ii) finetuning on BilPeD-Interview with an initialization using weights that are learned on FID. Notice that the sample size of BilPeD-Interview is significantly lower than that of FID.

As indicated in Section IV-A, each session of BilPeD includes three videos recorded from different views. For the evaluation of body pose modality on BilPeD-Interview, we use the right-front side videos, yet, in all other experiments on BilPeD, frontal videos are employed.

1) *Facial Appearance*: As described in Section III-A, we employ two different architectures, i.e., 3D ResNext-101 (3D-ResNext) and CNN-GRU, for modeling facial appearance. Both of these models are evaluated on BilPeD-Interview and FID, and obtained MAE results are given in Table III. On FID, CNN-GRU architecture provides an average MAE of 0.0101, reaching the baseline results provided in [16]. On the other hand, 3D ResNext-101 provides the most promising results on

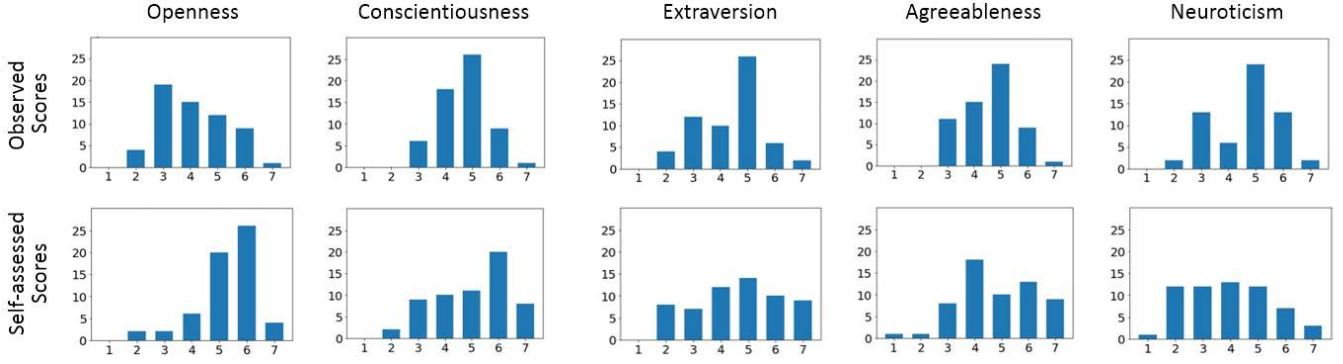


Fig. 8. Histograms of the observed and self-assessed scores for the five personality traits of 60 participants in BilPeD.

 TABLE IV
 MEAN ABSOLUTE ERRORS FOR THE USE OF FACIAL ACTION UNITS AND HEAD POSE & GAZE ON FID.

Model	Features	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	Facial AU	0.103	0.117	0.110	0.107	0.108	0.108
	Head Pose & Gaze	0.107	0.125	0.122	0.123	0.119	0.119
RCNN	Facial AU	0.102	0.116	0.102	0.109	0.104	0.106
	Head Pose & Gaze	0.109	0.127	0.124	0.124	0.119	0.121

FID with an average MAE of 0.088. Notice that the state-of-the-art method [16] (employing facial appearance, voice, and transcribed speech) on FID provides an MAE of 0.083, where the visual baseline MAE in [16] is only 0.096. Success of the 3D ResNext-101 on FID may rely on the 3D temporal convolutions.

While the lowest average MAE (0.156) on BilPeD-Interview is achieved using pretrained version of CNN-GRU, its MAE without pretraining is 1.3% (absolute) higher than that of 3D ResNext-101. Our results also show that pretraining on FID for BilPeD-Interview is useful although the structure of databases are different. Notice that FID has recordings extracted mostly from YouTube video blogs, while BilPeD-Interview includes recordings of answers to some questions. This finding can be explained by the fact that BilPeD-Interview is a relatively small database. Lastly, obtained MAEs for BilPeD are significantly higher than those of FID. This may suggest that during answering questions, facial cues of personality would be less visible compared to expression characteristics displayed in video blogs.

2) *Facial Action Units and Head Pose & Gaze*: In this experiment, we first assess the reliability of using facial action units and head pose & gaze on FID. Mean absolute errors of LSTNet and RCNN models on FID are given in Table IV. As shown by the results, MAEs for the use of action units through LSTNet and RCNN models are 0.108 and 0.106, respectively. Yet, the use of head pose & gaze performs 13% (absolute) worse on average than using facial action units. Consequently, on BilPeD-Interview we evaluate only the use of action units with an without a pretraining on FID.

 TABLE V
 MEAN ABSOLUTE ERRORS FOR THE USE OF FACIAL ACTION UNITS ON BILPED-INTERVIEW. * DENOTES PRETRAINING ON FID.

Model	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	0.170	0.163	0.205	0.188	0.194	0.184
LSTNet*	0.18	0.155	0.190	0.208	0.221	0.190
RCNN	0.221	0.205	0.220	0.216	0.218	0.216
RCNN*	0.149	0.125	0.172	0.174	0.183	0.161

 TABLE VI
 MEAN ABSOLUTE ERRORS FOR THE USE OF BODY POSE ON BILPED-INTERVIEW.

Model	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	0.159	0.168	0.182	0.202	0.178	0.178

As shown in Table V, the lowest MAE (0.161) on BilPeD-Interview using action units is obtained through RCNN model with a pretraining on FID. On the other hand, LSTNet architecture provides a 3.2% (absolute) lower MAE compared to RCNN if a pretraining on FID is not employed. Interestingly, enabling pretraining on FID reduces the performance of LSTNet by 0.6% (absolute) for this task.

3) *Body Pose*: For the evaluation of using body pose for estimating personality traits, only BilPeD-Interview is employed since the videos in FID do not show the whole body of subjects. To this end, we use the videos recorded from the right-front side in our experiment. Table VI shows that the use of body pose performs slightly better than random guess when used alone. Yet, with a large amount of data and powerful fusion strategies, body pose information would be expected to be useful.

4) *Voice*: The use of voice for modeling personality traits through LSTNet is evaluated on FID and BilPeD-Interview. As shown in Table VII, an average MAE of 0.12 is achieved on FID. On BilPeD-Interview, the obtained MAEs are lower than that of using facial action units and body pose, yet, significantly higher compared to the MAE on FID. Lastly, the LSTNet model pretrained on FID performs better for the voice modality.

TABLE VII
MEAN ABSOLUTE ERRORS FOR THE USE OF VOICE. * DENOTES
PRETRAINING ON FID.

Database	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	LSTNet	0.108	0.126	0.123	0.124	0.118	0.120
BilPeD-	LSTNet	0.176	0.141	0.200	0.177	0.196	0.178
Interview	LSTNet*	0.153	0.139	0.172	0.184	0.184	0.167

TABLE VIII
MEAN ABSOLUTE ERRORS FOR THE USE OF TRANSCRIBED SPEECH.
* DENOTES PRETRAINING ON FID.

Dataset	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	LSTNet	0.103	0.117	0.121	0.117	0.113	0.114
BilPeD-	LSTNet	0.158	0.150	0.189	0.184	0.198	0.176
Interview	LSTNet*	0.153	0.135	0.168	0.169	0.176	0.160

5) *Transcribed Speech*: In this experiment, we assess the discriminative power of transcribed speech for estimating personality traits. For a fair comparison, the speech in both BilPeD-Interview and FID videos are automatically transcribed using Google's Speech to Text API [28] as indicated in Section III-E. Next, the extracted transcriptions are used for language processing. Since, BilPeD-Interview has recordings in two languages, i.e., English and Turkish, multilingual embedding models are used in our experiment (both for BilPeD and FID).

As shown in Table VIII, the use of transcribed speech performs with an MAE of 0.114 on FID. MAEs obtained on BilPeD-Interview are 0.160 and 0.176, with and without a pretraining on FID, respectively. One of the reasons behind the MAE gap between FID and BilPeD-Interview may be due to the difference between the durations of their recordings. While each recording in FID is about 15 seconds, BilPeD-Interview has samples of about 60 seconds.

D. Combined Use of Modalities

In this study, we systematically evaluate the reliability of several modalities for the estimation of personality traits. In this experiment, we aim to find the most informative set of modalities through a high-level fusion strategy. To this end, we concatenate the predicted (regression) scores of the big five traits for each of the modalities to be included in the fusion. These score vectors are then modeled by a Linear Support Vector Machines Regressor (LSVR). For estimating each of the five traits, a separate LSVR model is employed. In this manner, the use of all possible combinations of facial appearance, facial action unit, body pose (only for BilPeD-Interview), voice, and transcribed speech modalities are evaluated in terms of MAE. To obtain the initial/unimodal personality scores, the following models are used: 3D-ResNext and CNN-GRU for facial appearance, LSTNet for facial action units, voice, transcribed speech, and body pose. For training models on BilPeD-Interview, FID pretraining is utilized except for body

pose modality. This is due to the fact that FID videos display only the upper body. Next, based on the resulting average MAE scores, we find the best performing (providing the minimum MAE) set of modalities on each of FID and BilPeD.

Our results show that the combination of facial appearance with CNN-GRU, facial appearance with 3D-ResNext, action units, and transcribed speech performs best on FID with an average MAE of 0.085. This is a 3.4% relative improvement over the sole use of the best performing modality, i.e., facial appearance (3D-ResNext) on FID. The worst performance on FID is observed for the combined use of voice and transcribed speech with an average MAE of 0.114. On BilPeD-Interview, the most reliable combination is found as facial appearance with CNN-GRU and transcribed speech, providing an average MAE of 0.149. This means a 4.5% relative improvement over the sole use of the best performing modality, i.e., facial appearance (CNN-GRU) on BilPeD-Interview. Interestingly, the combination of all modalities perform worst on BilPeD-Interview (with an average MAE of 0.172).

E. Analysis of Induced Behavior

Assessment of personality traits from induced behavior is one of the main goals of this study. To this end, as described in Section IV-A we have recorded 60 subjects while they are watching short video clips, referred to as BilPeD-Induction. Each of these video clips targets inducing the behavioral cues of (at least) one of the five personality traits. To this end, we employed five sets of video clips, namely for inducing openness, conscientiousness, extraversion, agreeableness, and neuroticism. In this way, we have formed five subsets in BilPeD-Induction, such as recordings during watching the aforementioned sets of inducing clips. Consequently, we aim to evaluate the use of induced behavior for estimating personality traits.

In the experiment, the sole use of facial appearance is utilized on BilPeD-Induction, since the results of our previous experiments suggest that the facial appearance is the most reliable modality for the analysis of personality. To this end, we employ both 3D ResNext-101 (3D-ResNext) and CNN-GRU models. Training of each model on BilPeD is initiated using weights that are learned on FID. A separate test is conducted for each of the five subsets (based on the induced traits).

As shown in Table IX, the minimum average MAE (0.153) is achieved using CNN-GRU. 3D-ResNext performs slightly worse with an average MAE of 0.155. Interestingly, except for conscientiousness and extraversion, the best performances are not achieved on the target (induced) traits. On the other hand, obtained results on BilPeD-Induction is slightly but consistently better than those on BilPeD-Interview (using facial appearance modality). This finding suggests that induced behavior through video clips indeed contains cues of personality traits. The most reliable predictions are obtained for conscientiousness both on FID and BilPeD-Interview, where the worst results are for openness. This may suggest that the visual cues of openness are more subtle and they a higher variance than those of other personality traits. Yet, further

TABLE IX

MEAN ABSOLUTE ERRORS FOR THE USE OF FACIAL APPEARANCE ON BILPED-INDUCTION. * DENOTES FINETUNING ON BILPED WITH AN INITIALIZATION USING WEIGHTS THAT ARE LEARNED ON FID.

Model	Induced Trait	MAE					
		AGR	CON	EXT	NEU	OPE	AVG
3D-ResNext*	AGR	0.145	0.113	0.168	0.166	0.184	0.155
	CON	0.147	0.117	0.171	0.167	0.179	0.156
	EXT	0.150	0.111	0.164	0.161	0.184	0.154
	NEU	0.151	0.122	0.165	0.157	0.172	0.154
	OPE	0.153	0.114	0.175	0.162	0.184	0.157
	AVG	0.149	0.115	0.169	0.163	0.181	0.155
CNN-GRU*	AGR	0.132	0.118	0.153	0.179	0.178	0.152
	CON	0.127	0.113	0.153	0.176	0.176	0.149
	EXT	0.136	0.117	0.153	0.169	0.179	0.151
	NEU	0.132	0.131	0.162	0.173	0.179	0.156
	OPE	0.144	0.117	0.171	0.180	0.182	0.159
	AVG	0.134	0.119	0.158	0.175	0.179	0.153

analysis is required for a more reliable interpretation of these findings.

VI. CONCLUSION

In this study, we have developed and presented spatio-temporal models that automatically assess the level of the Big Five personality traits (i.e., agreeableness, conscientiousness, extraversion, neuroticism, and openness), from different behavioral modalities such as facial appearance, facial action units, head pose & gaze, body pose, voice, and transcribed speech. State-of-the-art deep architectures have been utilized to this end. For a detailed analysis of personality, we have collected Bilkent Personality Database that consists of speech and induced behavior recordings of 60 participants. Each session in the database has three videos acquired from frontal, right-front side, and left-rear side cameras. In addition, Bilkent Personality Database includes both self-assessed and observed scores for each of the Big Five personality traits.

Using the developed methods, we have systematically evaluated the discriminative power of aforementioned modalities for the assessment of personality on two different databases, namely on Bilkent Personality and ChaLearn LAP First Impressions databases. In our experiments, facial appearance has been found to be the most reliable modality for personality analysis. The best performing combinations of behavioral modalities have also been analyzed. As well as providing baseline results on Bilkent Personality Database, we have reached the state-of-the-art results for ChaLearn LAP First Impressions Database. Based on our extensive experiments, we present new findings such that: (1) Induced behavior can display cues of personality; (2) People tend to express themselves as more socially desirable in their self-reports, when compared with the observed ones.

VII. ACKNOWLEDGEMENT

Authors would like to thank Mr. Alper Dağgez and Prof. Gül Günaydin for their contributions to this study.

REFERENCES

- [1] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.
- [2] L. R. Goldberg, "An alternative" description of personality": The Big-Five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.
- [3] L. M. P. Zillig, S. H. Hemenover, and R. A. Dienstbier, "What do we assess when we assess a big 5 trait? a content analysis of the affective, behavioral, and cognitive processes represented in big 5 personality inventories," *Personality and Social Psychology Bulletin*, vol. 28, no. 6, pp. 847–858, 2002.
- [4] M. S. Yik and J. A. Russell, "Predicting the big two of affect from the big five of personality," *Journal of Research in Personality*, vol. 35, no. 3, pp. 247–277, 2001.
- [5] A. A. Khan, K. C. Jacobson, C. O. Gardner, C. A. Prescott, and K. S. Kendler, "Personality and comorbidity of common psychiatric disorders," *The British Journal of Psychiatry*, vol. 186, no. 3, pp. 190–196, 2005.
- [6] H. Salem, A. Ruiz, S. Hernandez, K. Wahid, F. Cao, B. Karnes, S. Beasley, M. Sanches, E. Ashtari, and T. Pigott, "Borderline personality features in inpatients with bipolar disorder: Impact on course and machine learning model use to predict rapid readmission," *Journal of Psychiatric Practice*, vol. 25, no. 4, pp. 279–289, 2019.
- [7] N. Jakšić, L. Brajković, E. Ivezić, R. Topić, and M. Jakovljević, "The role of personality traits in posttraumatic stress disorder (ptsd)," *Psychiatria Danubina*, vol. 24, no. 3., pp. 256–266, 2012.
- [8] A. Cerasa, D. Lofaro, P. Cavedini, I. Martino, A. Bruni, A. Sarica, D. Mauro, G. Merante, I. Rossomanno, M. Rizzuto *et al.*, "Personality biomarkers of pathological gambling: A machine learning study," *Journal of neuroscience methods*, vol. 294, pp. 7–14, 2018.
- [9] J. M. Longenecker, R. F. Krueger, and S. R. Sponheim, "Personality traits across the psychosis spectrum: A hierarchical taxonomy of psychopathology conceptualization of clinical symptomatology," *Personality and mental health*, 2019.
- [10] J. N. Butcher, J. R. Graham, C. L. Williams, and Y. S. Ben-Porath, *Development and use of the MMPI-2 Content Scales*. University of Minnesota Press, 1990.
- [11] G. Cucurull, P. Rodríguez, V. O. Yazici, J. M. Gonfaus, F. X. Roca, and J. González, "Deep inference of personality traits by integrating image and word use in social networks," *arXiv preprint arXiv:1802.06757*, 2018.
- [12] C. Segalin, D. S. Cheng, and M. Cristani, "Social profiling through image understanding: Personality inference using convolutional neural networks," *Computer Vision and Image Understanding*, vol. 156, pp. 34–50, 2017.
- [13] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2017.
- [14] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 400–418.
- [15] F. Gürpınar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 43–48.
- [16] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baró, I. Guyon, J. J. Junior, M. Madadi *et al.*, "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," *arXiv preprint arXiv:1802.00745*, 2018.
- [17] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
- [18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1492–1500.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [20] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6546–6555.

- [21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [22] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using cnn-rnn and c3d hybrid networks,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.
- [23] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “Cnn-rnn: A unified framework for multi-label image classification,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [24] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, “Modeling long-and short-term temporal patterns with deep neural networks,” in *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2018, pp. 95–104.
- [25] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.
- [26] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.
- [27] T. Giannakopoulos, “pyaudioanalysis: An open-source python library for audio signal analysis,” *PloS one*, vol. 10, no. 12, 2015.
- [28] “Google cloud speech-to-text,” <https://cloud.google.com/speech-to-text>, accessed: 2019-09-15.
- [29] Huggingface, “huggingface/pytorch-transformers,” Jul 2019. [Online]. Available: <https://github.com/huggingface/pytorch-transformers>
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [31] D. J. Sparkman, S. Eidelman, A. R. Dueweke, M. S. Marin, and B. Dominguez, “Open to diversity,” *Journal of Individual Differences*, 2018.
- [32] S. Hassan, N. Akhtar, and A. K. Yılmaz, “Impact of the conscientiousness as personality trait on both job and organizational performance,” *Journal of Managerial Sciences*, vol. 10, no. 1, 2016.
- [33] H. R. Riggio and R. E. Riggio, “Emotional expressiveness, extraversion, and neuroticism: A meta-analysis,” *Journal of Nonverbal Behavior*, vol. 26, no. 4, pp. 195–218, 2002.
- [34] W. G. Graziano and N. Eisenberg, “Agreeableness: A dimension of personality,” in *Handbook of Personality Psychology*. Elsevier, 1997, pp. 795–824.



Selim Fırat Yılmaz was Born in Ankara, Turkey. He received his B.Sc. at Bilkent University Computer Engineering Department, in August 2019. He is currently a research assistant and an M.Sc. student at Bilkent University Electrical and Electronics Engineering Department, under the supervision of Prof. Suleyman S. Kozat. His research interests include online learning, anomaly detection, and natural language processing.



Can Ufuk Ertenli was born in Ankara, Turkey in 1995. He received a double-major B.Sc. in Electrical & Electronics Engineering and Mathematics from Middle East Technical University (METU), in 2018 and 2019, respectively. He is currently a research assistant and an M.Sc. student at METU Computer Engineering Department. His research interests include deep learning and computer vision.



Berhan Faruk Akgür received his psychology undergraduate degree from Bilkent University in 2017. His undergraduate thesis titled Effects of Illegal Substance Usage and Abuse on Comprehension, Metaphorical Meanings Interpretation, Personality Traits, and Evoked Emotions. He is currently pursuing his Ph.D. at Bilkent University Neuroscience program and working in National Magnetic Resonance Research Center (UMRAM) as a researcher. His research is about inhibitory-excitatory center & surround visual receptive fields.



Merve Kınıkloğlu received her B.Sc. degree in Psychology from TOBB University of Economics and Technology in August 2018. She is currently a Ph.D. student and teaching assistant in Neuroscience Program at Bilkent University working on size-contrast interaction in motion perception. Her research interest are antagonistic receptive field properties in visual cortex, visual perception and motion coherence.



Burak Mandıra received his B.Sc. degree in Computer Engineering at Bilkent University, Ankara, Turkey, in 2018. After graduation, he started his M.S. degree in Computer Engineering. He is currently a teaching assistant and M.Sc. student at Bilkent University. He is working on his thesis under the supervision of Dr. Hamdi Dibeklioğlu. His research interests include computer vision, deep learning, machine learning and neural networks.



Ash Gül Kurt received her B.Sc. degree from Psychology Department, Bilkent University, in Ankara, Turkey. She is currently a research assistant and a M.S. student in Neuroscience Department, Bilkent University, also known as UMRAM (National Magnetic Resonance Research Center). In her M.S. degree, she mainly focuses on motion discrimination and center-surround antagonism. Apart from visual perception, she is also interested in topics as personality, emotions, addiction, and UX.



Dersu Giritlioğlu was born in Ankara, Turkey in 1995. He received his B.Sc. degree at Middle East Technical University (METU) Electrical and Electronics Engineering Department, in June 2018. He is currently a teaching assistant and a M.Sc. student at I.D. Bilkent University. His research interest is in the intersection of computer vision, deep learning and behavioral psychology. He works on his thesis with Dr. Hamdi Dibeklioğlu on Audio-Visual Behavioral Analysis of Conversations.



Merve Nur Doğanlı is an undergraduate student at Ankara Yıldırım Beyazıt University, in the Department of Psychology. Her research interests in the field are particularly at the intersection of cognition and personality.



Emre Mutlu was born in Amasya, Turkey, in 1988. He graduated from the Ankara University Faculty of Medicine, in 2013 and completed his psychiatry training at the Hacettepe University Department of Psychiatry, in 2018. He is currently working as a psychiatry specialist at Etimesgut Şehit Sait Ertürk State Hospital, in Ankara. His research interests include personality traits, psychotic disorders and cognitive deficits in psychiatric disorders.



Şeref Can Gürel was born in London UK in 1980. He graduated from Hacettepe University Faculty of Medicine in 2004. He finished his psychiatry specialist training in the Department of Psychiatry Hacettepe University Ankara, later in 2013 he started to work in the same department as an assistant professor. Between 2016 and 2017 he worked under the supervision of professor Alexander T. Sack in Maastricht University Brain Stimulation and Cognition group on visual attentional orienting using methods of transcranial magnetic stimulation and

functional MRI. His current research interests are neuro-stimulation including ECT, TMS and treatment resistance in mental disorders.



Hamdi Dibeklioglu is an Assistant Professor in the Computer Engineering Department of Bilkent University, Ankara, Turkey, as well as being a Research Affiliate with the Pattern Recognition & Bioinformatics Group of Delft University of Technology, Delft, the Netherlands. He received the Ph.D. degree from the University of Amsterdam, Amsterdam, the Netherlands, in 2014. Before joining Bilkent University, he was a Postdoctoral Researcher at Delft University of Technology. His research focuses on Computer Vision, Pattern Recognition, Affective Computing, and Computer Analysis of Human Behavior. Dr. Dibeklioglu is a Program Committee Member for several top tier conferences in these areas. He was a Co-chair for the Netherlands Conference on Computer Vision 2015, a Local Arrangements Co-chair for the European Conference on Computer Vision 2016, a Publication Co-chair for the European Conference on Computer Vision 2018, and a Co-chair for the eNTERFACE Workshop on Multimodal Interfaces 2019.

Preliminary Results in Evaluating the Pleasantness of an Interviewing Candidate Based on Psychophysiological Signals

Didem Gökçay ⁽¹⁾, Fikret Arı ⁽²⁾, Bilgin Avenoglu ⁽³⁾, Fatih İleri ⁽¹⁾, Ekin Can Erkuş ⁽⁴⁾, Merve Balık ⁽⁵⁾, Anıl B. Delikaya ⁽¹⁾, Atıl İlerialkan ⁽¹⁾, Hüseyin Hacıhabiboğlu ⁽¹⁾

⁽¹⁾ Informatics Institute, Middle East Technical University, Ankara, Turkey

⁽²⁾ Department of Electrical Engineering, Ankara University, Ankara, Turkey

⁽³⁾ Department of Computer Engineering, TED University, Ankara, Turkey

⁽⁴⁾ Department of Biomedical Engineering, Middle East Technical University, Ankara, Turkey

⁽⁵⁾ Department of Psychology, Middle East Technical University, Ankara, Turkey

didemgokcay@gmail.com, Fikret.Ari@eng.ankara.edu.tr, bilgin.avenoglu@tedu.edu.tr, fatihileri@windowslive.com, eerkus@metu.edu.tr, merve.balik@metu.edu.tr, anilberkdelikaya@gmail.com, atililerialkan@gmail.com, hhuseyin@metu.edu.tr

Abstract— Job openings are getting scarce, as the number of applicants are increasing. Evaluation of job interviews is a serious burden for the human resource departments. Automatic evaluation of candidates applying for jobs is an important domain open for improvement. Evaluation of applicant videos from social signals is possible but may not be enough to determine candidates' suitability for jobs that require social stress. The interview itself is a social stressor, as widely known from the Trier social stress test. In this study we aimed to aid the candidate evaluation process, based on physiological signals collected from the facial area. More specifically, we tried to determine the pleasantness of video strips of an applicant based on the EMG, SCR, pupil and HR signals. A simple binary kNN classifier was able to determine the pleasantness with an accuracy of 73%.

Index Terms— Pleasantness, Pupil dilation, EMG, Skin conductance, Interview

I. INTRODUCTION

THE collaboration between humans and machines is increasing. Unfortunately, the potential role of social stress in human-machine interaction is not investigated much, even in

laboratory settings [1]. Social stress can be studied through three mechanisms: 1) The perceived stress by the human 2) The quantified stress of the human based on psychophysiological signals 3) The performance of the human [1]. Among these mechanisms, only the second one allows for automatic quantification of emotions during social stress. The Trier social stress test [2], which has been replicated reliably, focuses on the negative aspects of an interview situation, while several physiological measurements are collected from the participants invasively through blood samples and cortisol levels.

In this project, our objective is to collect and annotate multi-modal data in a laboratory setting that mimics an interview. Our main purpose is not to predict the stress, but to identify how pleasant the participant's interview is, considering the bodily responses during this stressful situation. Pleasantness and performance of a candidate could have been estimated through social signals or speech features as well. However, the scope of our study is limited with only psychophysiological features derived from pupil dilation, EMG, heart rate and skin conductance (SCR), mainly because we wanted to evaluate the potential of these measurements. Such automatic quantification might be beneficial for objective evaluation of the candidates and alleviate the workload of human resource management.

During stressful situations, the physiological changes initiated by the autonomous nervous system (ANS) are reviewed in detail in [3]. The pleasantness and arousal axes of the emotions are reflected differentially in body physiology [4]. More specifically, pupil dilation and skin conductance are linked to arousal, and EMG and heart rate are linked to valence (i.e. pleasantness). However there also exists a mild quadratic relationship between arousal and valence [4].

In this study, we aim to develop specific hardware to collect physiological data from the facial area and annotate this data using specific software. Annotations will be carried out using simultaneously collected video and speech of the participants as ground truth for pleasantness. Annotated data is needed to be used as a guide in supervised classification of the pleasantness of the participant's interview. The proof of concept will be shown separately for each measurement via a simple supervised classifier, and fusion of the will be left to future work.

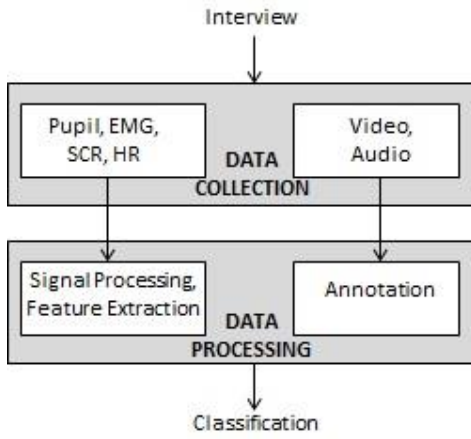


Fig. 1. Diagram of the candidate evaluation system during an interview

II. METHODS

The block diagram of the system used in this study is provided in Fig. 1. A mock interview is held with several legitimate questions, for which the participants' video, audio and bodily responses are recorded in the data collection phase. Then a data processing phase, consisting of three steps is run. First of all, the physiological signals are preprocessed to remove noise. Then, simultaneously captured video and audio are studied by two independent people to annotate pleasant and unpleasant segments. Finally, first level statistical features of each annotated segment are extracted separately from each physiological signal as features to be used in classification. At last, classification is done. A separate supervised classifier is trained and tested for each bodily signal to identify whether the interview was pleasant or not.

A. Data Collection

1) Hardware

In order to collect data from several sensors (infrared camera for pupil dilation, electromyography sensors for muscle movement, skin conductance response for sweating, heart beat sensor for heart rate) we used two electronic development boards. Synchronization is not a trivial issue. We had two problems with synchronization. First, the boards should be set with the same time stamp when the data collection begins. Second, time drift should be avoided. Drift is possible when the clocks of the boards are not executed in a synchronized manner especially for long periods of time. According to a study [5], almost 600 ms time drift is possible for a period of ten seconds. To avoid these problems we used two Raspberry Pi boards with a mutual clock. One board collected physiology data, the other collected video and audio data from a single HDMI camera. The time resolution of the sensors are as follows: Pupil diameter: 33 Hz, EMG (envelope) 30 Hz, SCR: 50 Hz, HR: 1.66 Hz. Sensor placement is as shown in Fig.2.

2) Participants

'Candidates' are selected among friends and volunteering colleagues attending the eNTERFACE '19 workshop. Only very limited pilot data is collected to solidify the experimental design for an ethical board application in the future.



Fig. 2. Wearable hardware components to collect psychophysiological changes while the participant answers interview questions

3) Procedure

A mock interview is held with 5 basic questions that are asked routinely in a typical job application interview (eg. 'Please introduce yourself', 'What position would you like to hold in this company within the next 5 years?', 'What important characteristics do you have, which makes you stand out from the crowd of other applicants?'... etc). When the candidates arrived, they are administered surveys regarding their current mood (PANAS), their depression level (BDI), and their anxiety level (STAI). After these surveys, the participants received the questions in print, and given time to think about their answers. The participants are allowed to take notes about their answers on note cards. Then they were fitted with the wearable sensors and camera. The experiment lasted approximately 20 minutes. During the debriefing, the participants were questioned informally about their stress level and how they performed.

B. Data processing

Signal processing and annotation are done on separate data streams, one on physiological data, the other on video/audio data.

1) Signal Processing and Feature Extraction

Each signal is preprocessed separately for noise removal and normalization. The pupil data is processed according to the steps described in [6], which involved interpolation for eye-blinks, as well as moving average and moving median filters for smoothing. The EMG data is not preprocessed, because the sensors returned the envelope of the data, which was already preprocessed. Each participant has a different physiological standing, which results in a different baseline in the associated signal. In order to train a classifier, such inter-personal signal variances must be normalized. For this purpose, we mapped each participant's signal to a range of 0-1 based on the entire series of responses recorded from the interview. Wherever the maximum is, that value is mapped to 1, wherever the minimum is, that value is mapped to 0. And the rest of the values in that signal domain is mapped between 0-1 linearly. Features are extracted separately for each signal simply by finding minimum, maximum and mean values of each annotated segment. In Fig.3. sample data from all signals during the recorded response for a single interview question are shown, such that the annotated segment is colored as orange.

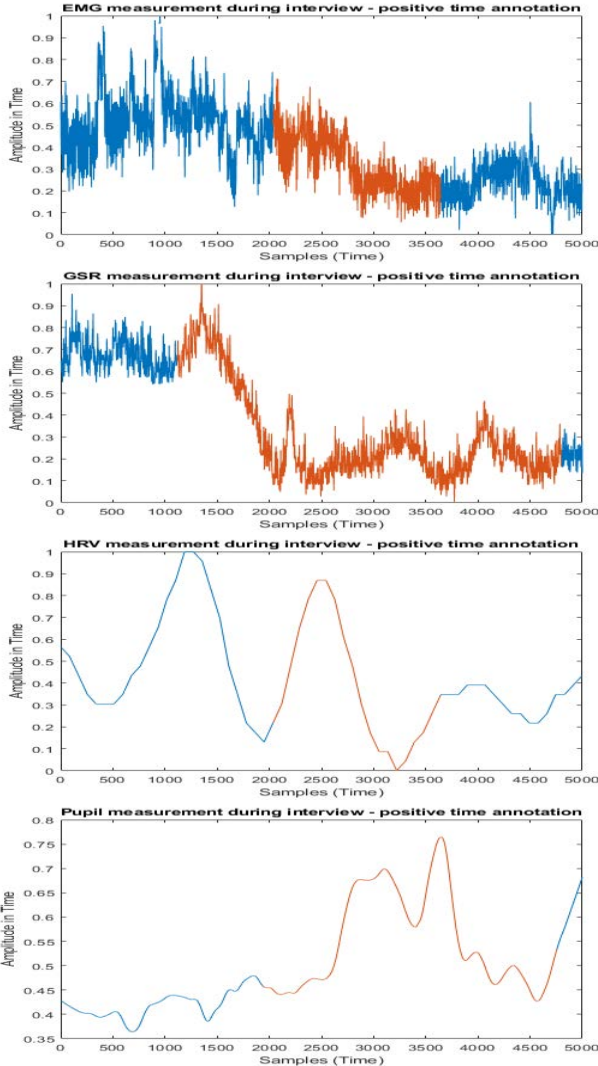


Fig. 3. A sample physiological signal set collected during one participant's response to an interview question. The signals from top to bottom are from EMG, SCR, HR, Pupil data. Orange regions identify annotated segments.

2) Annotation

We developed software to allow an impartial person to watch the videos, and tag the beginning and end of a segment. There are 4 different types of tags (each having *beginning* and *end* options): enthusiasm, fluency, stress, boredom. The tags were not mutually exclusive; they could be written on overlapping parts of the video stream. Enthusiasm and fluency characterized positive (i.e. pleasant) performance while stress and boredom characterized negative (i.e. unpleasant) performance. Tags are only written if the person who is making the decision had no doubt about its category. An entire video recording is allowed to be untagged if no segments are relevant.

C. Classification

Each annotated data segment is represented with a feature vector with 3 components (min, max and mean values of the associated signal in that segment) and a binary class: pleasant or unpleasant. We use the simplest method available in MATLAB, not because of preference, but due to simplicity. Two kNN classifiers are tested with $k=3$, $k=7$. The dataset is divided into 10 bins to allow for 10-fold cross validation.

TABLE I
DEMOGRAPHIC INFORMATION OF THE PARTICIPANTS (N=9)

	Mean \pm SD
Age	26 \pm 5.43
BDI	5.55 \pm 4.66
PANAS(+)	48.44 \pm 2.99
PANAS(-)	29.55 \pm 5.06
STAI(S)	32.33 \pm 8.47
STAI(T)	41.22 \pm 4.71

TABLE II
CLASSIFICATION ACCURACY OF K-NN FOR PLEASANTNESS

	K=3	K=7
EMG	60.87	73.91
SCR	69.57	73.91
HR	69.57	69.57
Pupil	73.91	73.91

III. RESULTS

There are 9 subjects (2 F, 7 M). Subject demographics are listed in table 1. None of the subjects are excluded due to their mood, depression or anxiety level. One subject is excluded due to guessing the relationship with the experiment and stress. After de-briefing, all subjects indicated that they were stressed during the interview, simply because of the gadgets and the pressure for performing well. Furthermore, the average STAI anxiety test results confirmed this. The STAI scores were at the high-end of the natural anxiety scale.

As seen in table 2, the accuracy performances of the 3 neighbor and 7 neighbor k-NN are close, varying between 69-74%. Unfortunately fusion of these signals even through a simple likelihood scheme using classifier level fusion could not be implemented due to the limited time allotted for this study.

IV. DISCUSSION

There exists several databases built on measuring affective responses from Visual, Audio, Eye Gaze, ECG, GSR, Respiration Amplitude, Skin temperature, and EEG [7] on a multitude of tasks. However, data sets consisting of multi-modal responses during social stress are scarce[1]. A well known social stress experiment, the Trier social stress test [2] involves an interview scenario. In this study, we used a mock interview setting and attempted to develop data collection and data processing pipelines for capturing physiological signals and analysing their contents for estimating pleasantness of the segments throughout the interview. This application has a potential for wide-spread use in the future because the job market is tightening, causing more and more candidates to apply for fewer open positions. Hence, automated scoring of interviews is around the corner.

While existing social signal processing and speech processing applications have a strong potential use in the video and audio signals collected during the interviews, we focused on a less investigated domain: the use of body physiology in the evaluation of interview performances. The results of the classification of individual signals with respect to pleasantness are on par with respect to other physiological signal

classification applications. The reported accuracy figures in the literature in tasks that involve stress vary widely between 60-95% accuracy, based on several factors such as restrictiveness of the environment, dynamic character of the task, invasiveness of the measurement devices, number of different measurements obtained from the body and smartness of the classification algorithms [8, 9, 10]. The task involved in this study is a semi-ambulant task, which allows freedom and dynamicity on the participant side, but restrictions on the experiment side due to the specific questions asked. The measurement devices collected multiple measurements, however, the sensors were invasive since they were not wireless and several body contacts existed. And the algorithm for classification was not a smart one, it did not even fuse the results through a simple classifier level fusion technique as suggested in [11]. Hence an accuracy on the order of 70-75% is promising, and warrants expectations of at least 10-15% increase if a few improvements are made.

V. LIMITATIONS AND FUTURE WORK

As suggested by the reviews [8, 9, 10], fusion methods can be used to improve the accuracy of binary prediction. Avoiding the curse of dimensionality is crucial for both feature level and classifier level fusion. Therefore, a continuation of our study with more samples is imperative to be able to report classification accuracy in multi-modal data.

Other than the low number of samples, our study had several limitations contingent upon data collection. The hardware equipment used in this study is assembled during the eNTERFACE 2019 workshop and tested for the first time. We realized that the frame holding the gadgets is not stable. It gradually shifts downward as the interview proceeds. Furthermore, the recording conditions such as lighting of the experiment room and the positioning of the subjects should also be improved. In addition, the field of view of the video recordings was not the same for all subjects, introducing bias to the annotation stage. These are all trivial in comparison to the variability introduced by human factors such as the motivation and cooperation of the participants during a mock interview, which need to be addressed in the future as well.

VI. CONCLUSION

Professional companies that deliver hiring services look for several attributes (eg. adaptability, innovativeness, approachability, curiousness, integrity and ambition) within the content of the verbal responses of a good candidate [13]. On the other hand, video and audio of the speech of the candidate is also valuable because social signal analysis and prosody analysis can be conducted automatically. We investigated another set of outputs embodied in physiology, to decide whether pleasantness can be predicted from the facial sensors attached to the interviewing candidate. The results indicate that each individual psycho-physiological signal collected during an interview setting can be used to identify the emotional state of a candidate on a pleasantness axis with more or less similar accuracy (i.e. approximately 70-75%) as in any other application based on psychophysiology. It is obvious that a benchmark dataset is needed in this domain for guiding future applications for automatic candidate selection.

ACKNOWLEDGMENT

We thank Bilkent University and the organizers of the eNTERFACE 2019 workshop, Hamdi Dibeklioglu and Elif Sürer. In addition, we would like to acknowledge the valuable contributions of our ex-team member Dr. Selen Pehlivan, who had to leave due to relocation abroad.

REFERENCES

- [1] Sauer, A., Schmutz, S., Sonderegger, A., Messerli, N. 2019. Social stress and performance in human-machine interaction: a neglected research field, *Ergonomics*, *Ergonomics*, DOI: 10.1080/00140139.2019.1652353
- [2] Dickerson, S. S., and M. E. Kemeny. 2004. "Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research." *Psychological Bulletin*, 130(3): 355. doi:10.1037/0033-2909.130.3.355.
- [3] Sharma, N., and Gedeon, T., 2012. "Objective measures, sensors and computational techniques for stress recognition and classification: A survey.", *Computer methods and programs in biomedicine*, 1287-1301.
- [4] Gökçay, D. (2011). 'Emotional Axes: Psychology, Psychophysiology and Neuroanatomical Correlates', In: D. Gökçay, G. Yıldırım (Eds.), *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, IGI Global
- [5] Mani, S. K., Durairajan, R., Barford, P., & Sommers, J. (2018). A System for Clock Synchronization in an Internet of Things. Retrieved from <http://arxiv.org/abs/1806.02474>
- [6] Baltacı, S., Gökçay, D., 'Stress Detection in Human Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features', *International Journal of Human-Computer Interaction*, 2016, DOI: 10.1080/10447318.2016.1220069
- [7] Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I., AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups, *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, Scheduled to appear
- [8] Smets, E., DeRaedt W., Van Hoof, C., Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings, *IEEE Journal of Biomedical and Health Informatics* · November 2018, DOI: 10.1109/JBHI.2018.2883751
- [9] Alberdi, A., Aztiria, A., Basarab, A. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review, *Journal of Biomedical Informatics* 59: 49–75
- [10] Can, Y. S., Arnrich, B., Ersoy, C. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, *J of Biomedical Informatics*, Vol. 92, 2019, doi.org/10.1016/j.jbi.2019.103139
- [11] Gökçay, D., Eken, A., Baltacı, S., Binary Classification Using Neural and Clinical Features: An Application in Fibromyalgia with Likelihood based Decision Level Fusion, *IEEE Biomedical and Health Informatics*, 2019, doi:10.1109/JBHI.2018.2844300
- [12] Wagner, J., Lingens, F., Andre, E., & Kim, J., 2011. "Exploring fusion methods for multimodal emotion recognition with missing data.", *IEEE Transactions on Affective Computing* 2.4, 206-218.
- [13] Linked Talent solutions: The interview questions you should be asking: <https://business.linkedin.com/talent-solutions>



Didem Gökçay graduated from the Electrical and Electronics Department of the Middle East Technical University with BS and MS degrees. She was a Fulbright scholar at the Department of Computer and Information Science and Engineering Department of University of Florida, where she obtained her PhD. Later she became a postdoctoral researcher at University of California, San Diego. She is the director of the MetuNeuro Lab and works as an associate professor at the Department of Health Informatics at the Middle East Technical University, Informatics Institute. Her major field of study is cognitive and affective neuroscience. She uses

fMRI, fNIRS, thermal camera, EMG, skin conductance and pupillary responses for studying the interaction between cognition and emotion. Currently she is on leave.



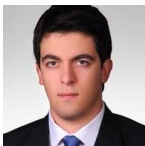
Fikret Arı received his B.S. degree in Electronics Engineering from Ankara University, Ankara, Turkey, in 1998. He received his M.S. and Ph.D. degrees in Electronics Engineering from Ankara University, Ankara, Turkey, in 2000 and 2006 respectively. Since December 1998, He has been with the Department of Electrical and Electronics Engineering at Ankara University, Ankara, Turkey, where he is currently an Associate Professor. His research interests include 1D/2D signal processing, optical communication and electromechanical equipment design.



Bilgin Avenoglu graduated from Gazi University with BS and Middle East Technical University with MS and PhD degrees in Information Systems. Currently he is Dr. Lecturer in TED University, Computer Engineering Department. His research interests include pervasive computing, intelligent environments, multi-agent systems and cognitive agents.



Fatih İleri took his Electrical and Electronics Engineering BS and MS degrees from Bahçeşehir University and Boğaziçi University, respectively. Currently, he is a senior software engineer in Turkish Aerospace Industries Inc. He is also a PhD student in Department of Medical Informatics in Middle East Technical University. His main research interests are pattern recognition, image and signal processing.



Ekin Can Erkuş took his Electrical and Electronics Engineering BS degree from Hacettepe University. He graduated from the Biomedical Engineering department of Middle East University. Currently, he is a PhD student at the Department of Biomedical engineering in Middle East Technical University. His main research interests are biomedical signal analysis, pattern recognition, machine

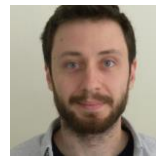
learning, data modelling, medical imaging and biocompatible devices.



Merve Balık is a junior student at the Department of Psychology in Middle East Technical University. Her main research interests are cognitive neuroscience, brain disorders, mental dysfunction.



Anıl Berk Delikaya took his BS degree in psychology from Hacettepe University. Currently, he is an MS student at the Department of Medical Informatics in Middle East Technical University.




Atıl İlerialkan took his BS degree in Computer Engineering from Hacettepe University and MS degree in Multimedia Computing from the Informatics Institute of Middle East Technical University. Currently, he works as a senior software engineer at Turkish Aerospace Industries. His main research interests are speech processing and machine learning.



Hüseyin Hacıhabiboğlu is an Associate Professor of Signal Processing at the Graduate School of Informatics, Middle East Technical University (METU). He received the BSc degree from METU in 2000, the MSc degree from the University of Bristol in 2001, both in electrical and electronic engineering, and the PhD degree in computer science from Queen's University Belfast. He held research positions at University of Surrey and King's College London. His research interests include audio signal processing, room acoustics, multichannel audio systems, psychoacoustics of spatial hearing, microphone arrays, and game audio. He is a Senior Member of the IEEE, Member of the IEEE Signal Processing Society, Audio Engineering Society (AES), Turkish Acoustics Society (TAD), and the European Acoustics Association (EAA) and an associate editor of *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Volleyball Action Modelling for Behavior Analysis and Interactive Multi-modal Feedback

Fahim A. Salim ⁽¹⁾, Fasih Haider ⁽²⁾ , Sena Busra Yengec Tasdemir ⁽³⁾, Vahid Naghashi ⁽⁴⁾, Izem Tengiz ⁽⁵⁾, Kubra Cengiz ⁽⁶⁾, Dees B.W. Postma ⁽⁷⁾, Robby van Delden ⁽⁷⁾, Dennis Reidsma ⁽⁷⁾, Saturnino Luz ⁽²⁾, Bert-Jan van Beijnum ⁽¹⁾

⁽¹⁾ Biomedical Signals and Systems, University of Twente, The Netherlands

⁽²⁾ Usher Institute, Edinburgh Medical School, the University of Edinburgh, United Kingdom

⁽³⁾ Electrical Computer Engineering Department, Abdullah Gul University, Turkey

⁽⁴⁾ Computer Engineering Department, Bilkent University, Turkey

⁽⁵⁾ Department of Biomedical Engineering, Izmir University of Economics, Turkey

⁽⁶⁾ Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

⁽⁷⁾ Human Media Interaction, University of Twente, The Netherlands

f.a.salim@utwente.nl, Fasih.Haider@ed.ac.uk

Abstract—Quick and easy access to performance data during matches and training sessions is important for both players and coaches. While there are many video tagging systems available, these systems require manual efforts. In this project, we use Inertial Measurement Units (IMU) sensors strapped on the wrists of volleyball players to capture motion data and use Machine Learning techniques to model their actions and non-actions events during matches and training sessions.

Analysis of the results suggests that all sensors in the IMU (i.e. magnetometer, accelerometer, barometer and gyroscope) contribute unique information in the classification of volleyball-specific actions. We demonstrate that while the accelerometer feature set provides the best Unweighted Average Recall (UAR) overall, decision fusion of the accelerometer with the magnetometer improves UAR slightly from 85.86% to 86.9%. Interestingly, it is also demonstrated that the non-dominant hand provides better UAR than the dominant hand. These results are even more marked with decision fusion.

Apart from machine learning models, the project proposes a modular architecture for a system to automatically supplement video recording by detecting events of interests in volleyball matches and training sessions and to provide tailored and interactive multi-modal feedback by utilizing an html5/JavaScript application. A proof of concept prototype is also developed based on this architecture.

Index Terms—IEEE, IEEEtran, journal, L^AT_EX, paper, template.

I. INTRODUCTION

TOP performance in sports depends on training programs designed by team staff, with a regime of physical, technical, tactical and perceptual-cognitive exercises. Depending on how athletes perform, exercises are adapted, or the program may be redesigned. State of the art data science methods have led to ground breaking changes. Data is collected from sources such as tracking position and motion of athletes in basketball [1] and baseball and football match statistics [2].

Furthermore, new hardware platforms appear, such as LED displays integrated into a sports court [3] or custom tangible sports interfaces [4]. These offer possibilities for hybrid

training with a mix of technological and non-technological elements [3]. This has led to novel kinds of exercises [5], [4] including real-time feedback, that can be tailored to the specifics of athletes in a highly controlled way. Data science tools can then be used to precipitate tailored modifications to (the parameters of) such training.

These developments are not limited to elite sport. Interaction technologies are also used for youth sports (e.g., the widely used player development system of Dotcomsport.nl), and school sports and Physical Education [6].

This eNTERFACE project is a part of the Smart Sports Exercises (SSE) project which aims to extend the state of the art by combining sensor data, machine learning and interactive video to create new form of volleyball training and analysis.

For this particular project we focused on identifying volleyball actions performed by players by strapping IMUs (Inertial Measurement Unit) on their wrist(s) and using Machine Learning techniques to model and classify their actions. In addition to identifying the action, the second main aim of the project is to supplement the video recordings by automatically tagging (identify and provide a link to its timestamp) the identified action and events.

A. Motivation

Automatically identifying actions in sport activities is important for many reasons, therefore there have been numerous studies to identify actions in sports [7], [8], [9], [10]. Wearable devices such as Inertial Measurement Units (IMUs) [11], [12] are becoming increasingly popular for sports related action analysis because of their reasonable price as well as portability [10]. While researchers have proposed different configurations in terms of number and placement of sensors [13], it is ideal to keep the number of sensors to minimum due to issues related to cost, setup effort and player comfort [14], [15], [16], [13].

In addition to identification and analysis, access to performance data during sports matches and training sessions is

important for both players and coaches. Analysis of video recording showing different events of interest may help in getting insightful tactical play and engagement with players [17] and video edited game analysis is a common method for post-game performance evaluation [6].

Accessing events of interest in sports recording is of particular interest for both sports fans e.g. a baseball fan wishing to watch all home runs hit by their favorite player during the 2013 baseball season [7], or a coach searching for video recordings related to the intended learning focus for a player or the whole training session [6].

However, these examples require events to be manually tagged which not only requires time and effort but would also split a trainers attention from training to tagging the events for later viewing and analysis.

A system which could automatically tag such events would help trainers avoid manual effort has the potential to provide tailored and interactive multi-modal feedback to coaches and players.

B. Project Objectives

In summary, the project has the following objectives:

- To evaluate the potential of using sensor data from IMUs (3D acceleration, 3D angular velocity, 3D magnetometer and air pressure) in automatically identifying basic volleyball actions and non-action;
- to use Machine Learning techniques to identify individual player actions;
- to supplement the video recording by tagging the identified action and events, and
- to design a system to allow coaches and players to view tagged video footage to easily search for the information or event of interest (e.g. All the serves by a particular player)

II. RELATED WORK

Quick and easy access to performance data is important for both coaches and players, therefore it is important that video recordings related to the intended learning focus are immediately accessible [6]. In their work Koekoek et al. developed an application named Video Catch to manually tag events like sports actions during matches and training sessions [6]. Creating a system which can automatically tag such actions would be beneficial as it would save manual effort.

Inertial Measurement Units (IMUs) [11], [12] have been utilized to automatically detect sport activities in numerous sports e.g. soccer [18], [19], tennis [20], [21], table tennis [22], hockey [19], basketball [23], [24] and rugby [25].

Many approaches have been proposed for human activity recognition. They can be categorized into two main categories: sensor-based and vision-based.

Vision-based methods employ cameras to detect and recognize activities using several computer vision techniques. While sensor-based methods collect input signals from wearable sensors mounted on human bodies such as accelerometer and gyroscope. For example, In Liu et al. [26] identified temporal

patterns among actions and used those patterns to represent activities for the purpose of automated recognition. Kautz et al. [27] presented an automatic monitoring system for beach volleyball based on wearable sensor devices which are placed at wrist of dominant hand of players. Beach volleyball *serve* recognition from a wrist-worn gyroscope is proposed in Cuspinera et al. [28] which is placed on the forearm of players. Kos et al. [29] proposed a method for tennis stroke detection. They used a wearable IMU device which is located on the players wrists. A robust player segmentation algorithm and novel features are extracted from video frames, and finally, classification results for different classes of tennis strokes using Hidden Markov Model are reported [30]. Jarit et al. [31] studied with college baseball players. 88 subjects were studied in two groups. Jamar dynamometer was used to test maximum grip strength (kgf) for both hands. The recording was done for dominant and nondominant hands. The highest measurements were taken for the statistical analysis. Every subject put their maximal effort. 2-factor repeated measures to analyze the variance was used to compare both hands grip strength ratios of the experimental and control group. Results of the study showed that there is no significant differences of baseball players dominant and nondominant hands grip strength. Based on the above literature, we have concluded that the most studies take into account the role of dominant hand particularly for volleyball action modelling and the role of non-dominant hand is less explored.

III. METHODOLOGY

The project can be divided into following activities.

- Data Collection
- Prototype System
- Machine Learning (Feature Extraction and Modeling)

IV. DATA COLLECTION

A. Technical Setup

- Each player wears 2 IMUs (see Figure 1) on both wrists.
- Two video cameras on the side of team wearing the IMUs (see Figure 2).

B. Participants

Nine volleyball players wore IMU sensors [11] on both wrists during their regular training session. Players were encouraged to play normally as their routine training session. Due to some technical reasons IMUs wore by one player did not work, therefore the data used for the experimentation consists of 8 volleyball players.

C. Data Annotation

To obtain the ground truth for machine learning model training, the video recording was annotated using the Elan software (see Figure 3). 3 annotator annotated the video. Since volleyball actions performed by players are quite distinct there is no ambiguity in terms of inter-annotator agreement. The quality of the annotation is evaluated by a majority vote i.e. if



Fig. 1: Player wearing 2 IMUs on both wrists.

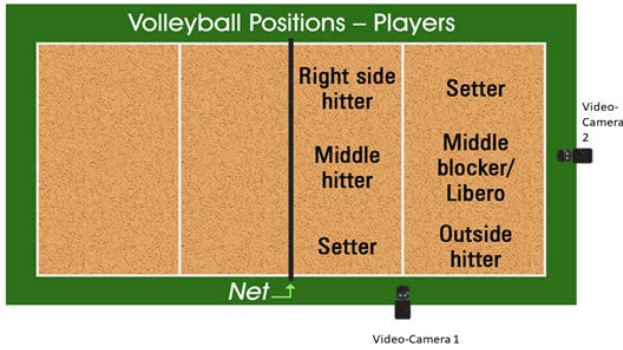


Fig. 2: Camera settings on court.

all annotator have annotated the same action or if an annotator might have missed or mislabelled an action.

As a result, for action case and non-action case there were 1453 and 24412 seconds of data, respectively. Table I shows the data (in seconds) information for each player. This data set is made available to research community. The annotators also annotated the type of volleyball actions such as under hand serve, overhead pass, serve, forearm pass, one hand pass, smash, underhand pass. Table I also details the number of volleyball actions performed by every player.

V. AUTO-TAGGING SYSTEM PROTOTYPE

The auto-tagging system has the following components.

A. Sensors on Player Wrist(s)

During a training session or a match, players wear a wireless sensor such as an IMU (Inertial Measurement Unit) [11], [12] on one or both wrists (see section IV for details). Features are extracted from the IMU signals to train machine learning models to recognize volleyball actions and non-actions. The machine learning is performed in two steps as shown in Figure 4, first we recognize if a frame of sensor data belongs

to a volley ball action or not. If it belongs to an action then we further classify it into types of actions (see VI for machine learning modelling and experimentation). Once the actions are identified, its information along with the timestamp is stored in a repository for indexing purposes.

B. Repository

Information related to the video, players and actions performed by the players are indexed and stored as documents in a tables or cores in Solr search platform [32]. An example of a Smash indexed by Solr is shown in table II.

C. Web Application

The interactive system is developed as web application. The server-side is written using asp.net MVC framework. While the front-end is developed using HTML5/Javascript.

Figure 5 shows a screen shot of the front-end of the developed system. The player list and actions list are dynamically populated by querying the repository. The viewer can filter the actions by player and action-type (e.g. over head pass by player 3). Once a particular action item is clicked or taped, the video is automatically jumped to the time interval where the action is being performed.

VI. EXPERIMENTAL SETUP

For classification, a two level task classification is planned. In the first step a binary classification scheme is adopted where the given frame (as described in section VI-A) is classified as Action or Non-Action. In the second step (future plan), the action in the window will be classified as Forearm Pass, One Hand Pass, Overhead Pass, Serve, Smash, Underhand Pass, Underhand Serve or Block. In this study, we have only trained machine learning models for action and non-action events (i.e. first step only). This section describes the process of machine learning models training for action and non-action events.

A. Feature Extraction

In this study, we have used time domain features such as mean, standard deviation, median, mode, skewness and kurtosis which are extracted over a frame length of 0.5 seconds of sensor data with an overlap of 50% with the neighbouring frame. As a results we have six features for each dimension of sensor data per frame. For action case and non-action case there were 5812 and 97648 frames, respectively.

B. Classification Methods

The classification is performed using five different methods namely Decision Tree (DT, with leaf size of 5), Nearest Neighbour (KNN with K=5), Naive Bayes (NB with kernel distribution assumption), Linear Discrimination Analysis (LDA) and Support Vector Machines (SVM with a linear kernel with box constraint of 0.5 and SMO solver). The classification methods are employed in both Python and MATLAB¹ using the statistics and machine learning toolbox in the Leave-One-Subject-Out (LOSO) cross-validation setting, where the

¹<http://uk.mathworks.com/products/matlab/> (December 2018)

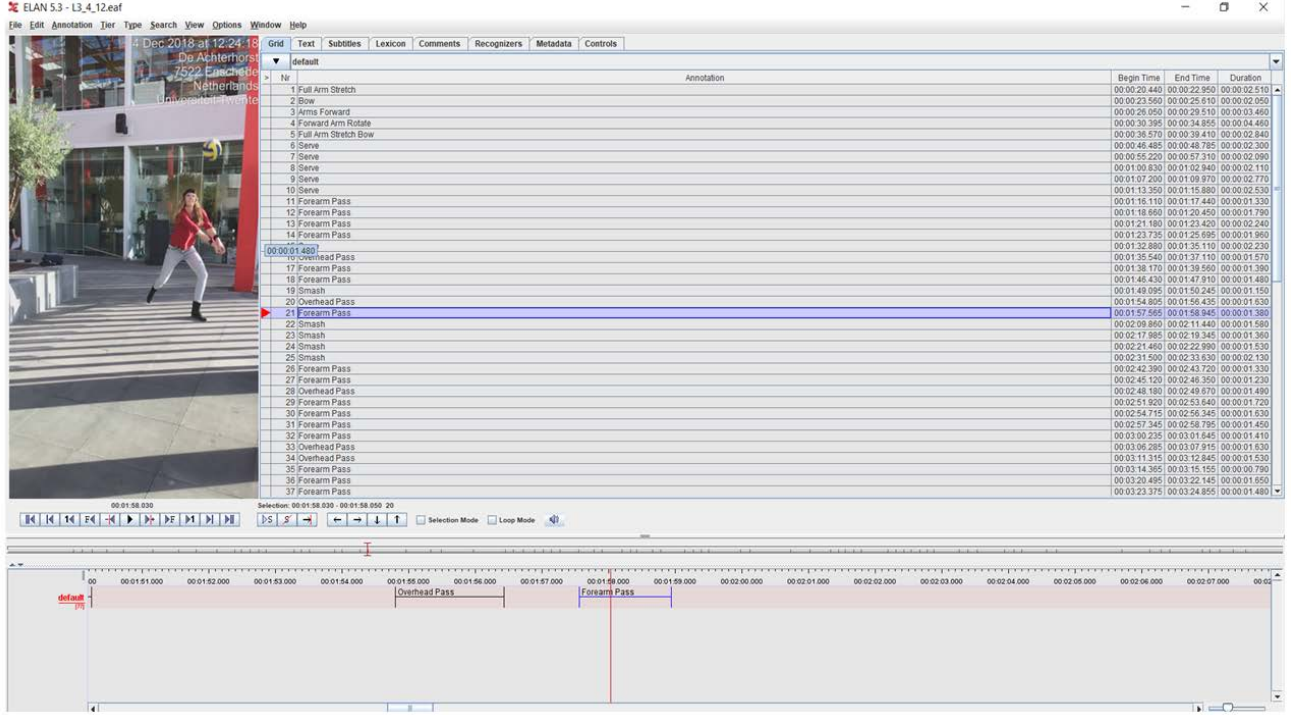


Fig. 3: Annotation example with Elan annotation tool.

TABLE I: Data Set Description: Time taken by each player for performing actions, non actions and number and type of actions performed by each player

ID	DH	Action(sec)	Non-Action(sec)	# Actions	Forearm Pass	Onehand Pass	Overhead Pass	Serve	Smash	Underhand Serve	Block
1	R	198	3055.25	120	40	3	16	0	29	28	4
2	L	193.75	3061	125	36	2	14	32	15	0	6
3	R	191	3030	116	50	3	3	34	25	0	1
5	R	176.75	3054.5	124	46	2	19	21	28	4	4
6	R	228.5	3009	150	30	1	70	0	12	30	7
7	R	135.5	3080.25	106	39	4	13	0	14	34	2
8	R	146.25	3077.5	105	34	4	16	34	17	0	0
9	R	183.25	3044.5	144	42	1	58	33	4	1	5
total		1453	24412	990	317	20	209	154	144	97	49

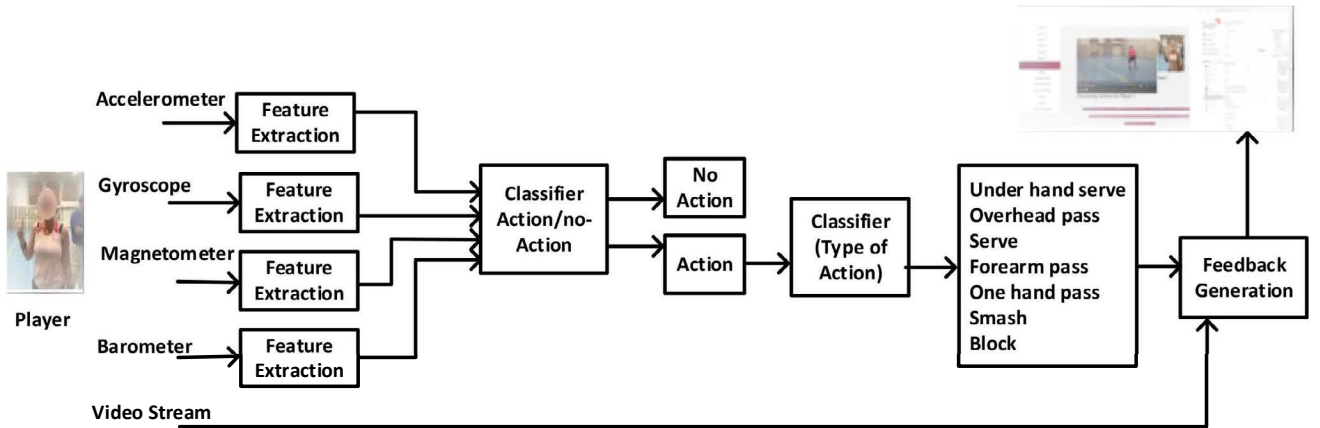


Fig. 4: Prototype System Architecture

training data do not contain any information of validation subjects. To assess the classification results, we used the unweighted average recall as the dataset is not balanced. The unweighted average recall is the arithmetic average of recall

of both classes.

C. Experiments

The overall action frames for eight players were 5812 frames while in Non-Action case there were 97648 frames.

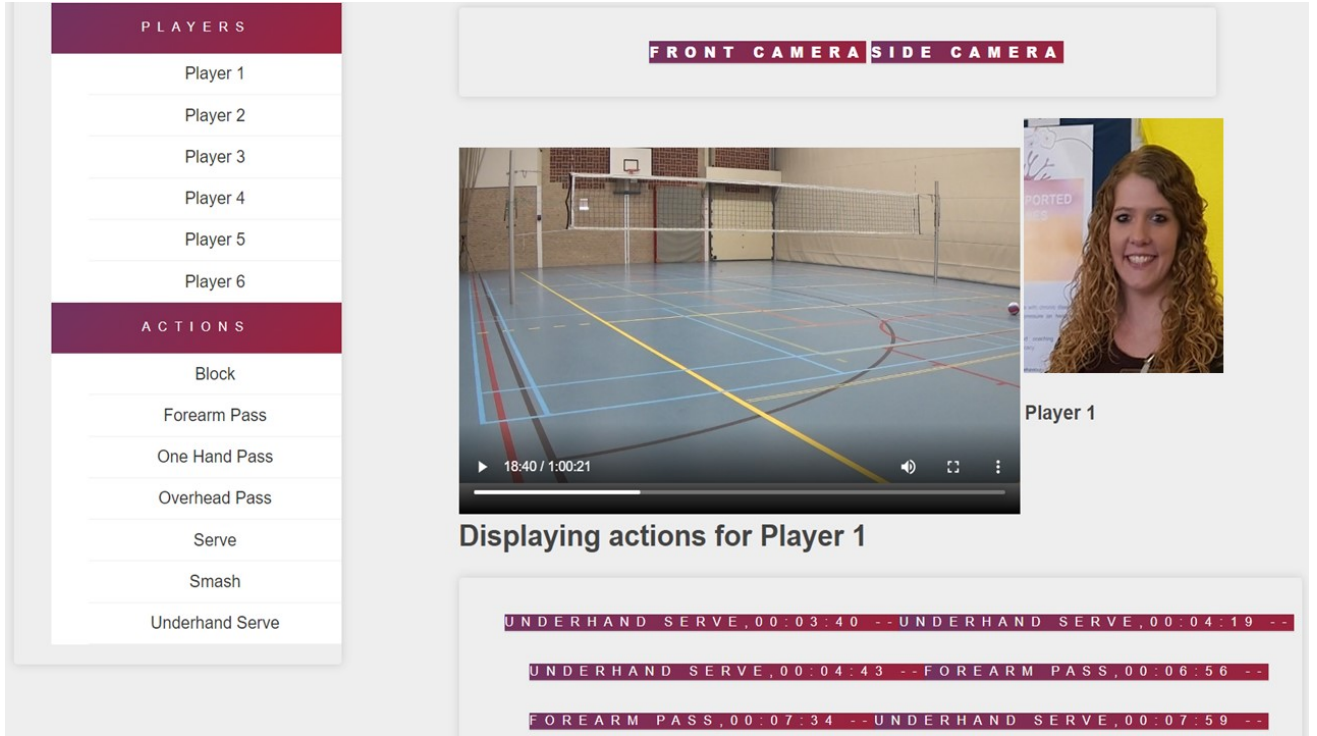


Fig. 5: Interactive front-end system

TABLE II: Sample Solr structure

```
"id":"25_06_Player_1_action_2"
"player_id":["25_06_Player_1"],
"action_name":["Smash"],
"timestamp":["00:02:15"],
"_version_":1638860511128846336
```

One can understand from the samples that the data set is imbalanced. In order to evaluate the performance of IMU sensor, we train machine learning models using balanced and imbalanced data set for the recognition of Action and non Action frames, we have conducted two experiments as follow:

- **Experiment 1:** training is performed on balanced data sets in terms of actions and non actions, where the same number of non-actions events (selected randomly) and action events for each player are used. The validation is performed on imbalanced (full) dataset in leave-one-subject out settings.
- **Experiment 2:** training is performed on imbalanced data sets in terms of action and non actions and validation is performed on imbalanced dataset in leave-one-subject out settings.

VII. EXPERIMENTAL RESULTS

This section describes the results of machine learning models for action and non-action events and demonstrate the discriminate power of different IMU sensors placed on dominant and non-dominant hand.

A. Experiment 1

The results of dominant hand and non-dominant hand for all sensors are shown in Table III and Table IV respectively. The best results indicate that the dominant hand (82.50%) provides better UAR than non-dominant hand (81.71%) using the accelerometer. The average of results indicated that the accelerometer provides the best averaged UAR of 81.92% (dominant hand) and 80.41% (non-dominate hand). SVM classifier provides the best averaged UAR of 74.36% (dominant hand) and 72.30% (non-dominate hand). All sensors provide better results (i.e. UAR) on dominant hand than on non-dominant hand.

TABLE III: Dominant Hand: Unweighted Average Recall (%)

Sensor	DT	KNN	NB	SVM	LDA	avg.
Acc.	81.99	82.50	82.19	82.35	80.52	81.91
Mag.	77.47	74.86	79.25	79.50	79.08	78.03
Gyr.	73.72	75.48	75.94	74.17	72.78	74.42
Baro.	57.19	56.80	59.30	61.45	61.01	59.15
avg.	72.59	72.41	74.17	74.36	73.34	—

TABLE IV: Non-Dominant Hand: Unweighted Average Recall (%)

Sensor	DT	KNN	NB	SVM	LDA	avg.
Acc.	78.90	80.33	81.71	81.28	79.84	80.41
Mag.	74.80	69.59	75.31	76.69	75.90	74.46
Gyr.	72.84	73.42	74.74	75.35	75.10	74.29
Baro.	51.57	50.22	49.46	55.88	56.07	52.64
avg.	69.52	68.39	70.30	72.30	71.72	—

B. Experiment 2

The UAR of dominant hand and non-dominant hand for all sensors are shown in Table V and Table VI respectively. These results indicate that the non-dominant hand (83.99%) provides better UAR than dominant hand (79.83%), with NB being the best classifier for action detection. The results indicated that the accelerometer provides the best averaged UAR of 69.76% (dominant hand) and 74.17% (non-dominant hand). NB classifier provides the best results of 71.47% (dominant hand) and 68.01% (non-dominant hand). The averaged UAR also indicates that the accelerometer (74.14%) and magnetometer (73.52%) provide better UAR on non-dominant hand than on dominant hand.

TABLE V: Dominant Hand: Unweighted Average Recall

Sensor	DT	KNN	NB	SVM	LDA	avg.
Acc.	70.83	68.83	79.83	59.77	69.56	69.76
Mag.	63.10	57.12	74.16	50.00	67.71	62.41
Gyr.	64.07	60.78	74.58	53.35	64.86	63.53
Baro.	59.22	56.53	57.24	53.01	56.78	56.56
avg.	64.30	60.81	71.45	54.03	64.72	—

TABLE VI: Non-Dominant Hand: Unweighted Average Recall

Sensor	DT	KNN	NB	SVM	LDA	avg.
Acc.	71.53	72.98	83.99	66.47	75.90	74.17
Mag.	76.61	67.67	80.83	66.75	75.74	73.52
Gyr.	61.42	58.85	75.71	50.00	64.70	62.14
Baro.	40.86	38.56	31.53	50.00	50.53	42.30
avg.	62.60	59.51	68.01	57.80	66.71	—

C. Sensor Fusion

We implemented a simple decision fusion strategy by taking a vote among all feature sets i.e fusing the output of the best classifiers for each sensor, breaking ties by considering them as implying a non-action label. The ‘fusion results’ of experiment 1 and experiment 2 are shown in Table VII and Table VIII respectively. The reported results are quite promising, indicating that the sensors placed on the wrist of players could be used to detect whether a player is performing a volleyball action or not. It also suggests that fusion of accelerometer and magnitude sensors provides the best results when placed on both hands. However placing magnitude and accelerometer on one hand provides slightly less accurate results than placing them on both hands. It is also observed that the fusion for Experiment 2 provides better results than fusion of Experiment 1. It could be due to the reason of training setup, as lesser data is used in experiment 1 than experiment 2. The average UAR of sensor fusion indicated that the fusion improves the UAR and the confusion matrix of best UAR for experiment 1 and experiment 2 are shown in Figure 6 and 7 respectively. This study will also help in lowering the number of sensors for the players which could results in cost reduction of system and making the system less intrusive.

The reported study is part of the Smart Sports Exercises project in which we aim to develop new forms of volleyball training using wearable sensors data and pressure sensitive in-floor displays to provide analysis and feedback in an interactive manner. While we are interested not only in action and

TABLE VII: Sensor Fusion: Unweighted Average Recall

Sensor	DH	NDH	Both Hands
acc	82.50	81.71	82.52
Mag	79.50	76.69	77.65
Gyr	75.94	75.35	76.42
Baro.	61.45	56.07	60.14
Acc + Mag	81.87	80.03	83.84
Acc + Gyr.	80.60	79.21	81.88
Gyr + Mag	78.64	76.97	81.08
Acc + Mag + Gyr	79.73	80.30	83.52
All	82.25	79.73	83.51
Avg.	78.05	76.82	78.95

TABLE VIII: Sensor Fusion: Unweighted Average Recall

Sensor	DH	NDH	Both Hands
acc	79.83	83.99	85.86
Mag	74.16	80.83	86.38
Gyr	74.58	75.71	81.25
Baro.	59.22	50.53	59.34
Acc + Mag	81.84	86.42	86.87
Acc + Gyr.	79.40	83.09	85.02
Gyr + Mag	78.34	82.91	86.08
Acc + Mag + Gyr	80.73	84.58	85.80
All	72.91	84.53	85.32
Avg.	75.66	79.17	82.43

		Precision	
		Non-Action	Action
Output Class	Non-Action	81635 78.9%	925 0.9%
	Action	16013 15.5%	4887 4.7%
Recall	Non-Action	83.6%	84.1%
	Action	16.4%	15.9%
		Accuracy	
		Non-Action	Action
		83.6%	83.6%
		16.4%	16.4%
		Target Class	
		Non-Action	Action

Fig. 6: Experiment 1 (Confusion matrix): best sensor fusion results obtained using fusion of accelerometer and magnetometer sensors from both hands.

non-action but also the type of action such as serve, forearm pass. It may be the case that dominant hand plays a crucially important role in determining the type of action. However, in many applications such as fatigue and stamina estimation [8], researchers are only interested in determining the amount of actions performed regardless of their type. In such cases, the reported results show an interesting case of using non-dominant hand compared to the common practice of using sensor(s) on the dominant hand [33], [34].

VIII. CONCLUSION

The overall aim of this project was to design an automatic video tagging system for sports related events using Machine Learning techniques and IMU sensors. In terms of contribution, this project proposed an architecture to automatically supplement video recordings, to this end; apart from the architecture, a prototype was developed based on that architecture

Output Class			Precision
	Non-Action	Action	
	Recall		Accuracy
Non-Action	82433 79.7%	621 0.6%	99.3% 0.7%
Action	15215 14.7%	5191 5.0%	25.4% 74.6%
Target Class	84.4% 15.6%	89.3% 10.7%	84.7% 15.3%

Fig. 7: Experiment 2 (Confusion matrix): best sensor fusion results obtained using fusion of accelerometer and magnetometer sensors from both hands.

as proof of concept. Secondly the project developed and tested machine learning models trained on IMU data.

The experimentation performed during the project provided interesting results not only in terms of UAR but also in terms of sensor configuration. The analysis of using non-dominant hand for sensor placement opened up interesting opportunities for sports research.

IX. FUTURE DIRECTIONS

The outcome of this eNTERFACE project has the potential to be extended in multiple ways. In terms of machine learning models, we aim to train models to not only classify action vs non-action but also type of volleyball actions such as under hand serve, overhead pass, serve, forearm pass, one hand pass, smash, underhand pass. Additionally we plan to use frequency domain features such as Scalogram and Spectrogram instead of time domain features currently used to train the models.

Apart from extending the machine learning models the aim is to further develop the video tagging system from a proof of concept prototype to a more functional and integrated system.

The following list summarises possible ways to extend the project.

- Further classify actions
- Using frequency domain approaches for feature extraction
- Scalogram, spectrogram
- ResNet, AlexNet, VGGNet
- Classification based on the above feature set.
- Further integration of Demo system and models.

ACKNOWLEDGMENT

This work is carried out as part of the Smart Sports Exercises project funded by ZonMw Netherlands and the European Union's Horizon 2020 research and innovation program, under the grant agreement No 769661, towards the SAAM project. Sena Busra Yengec Tasdemir, is supported by the Turkish Higher Education Council's 100/2000 PhD fellowship program.

REFERENCES

- [1] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: Current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017.
- [2] H. K. Stensland, Ø. Landsverk, C. Griwodz, P. Halvorsen, M. Stenhaug, D. Johansen, V. R. Gaddam, M. Tennøe, E. Helgedagsrud, M. Næss, H. K. Alstad, A. Mortensen, R. Langseth, and S. Ljødal, "Bagadus: An integrated real time system for soccer analytics," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 10, no. 1s, pp. 1–21, 2014. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2576908.2541011>
- [3] R. Kajastila, "Motion Games in Real Sports Environments," *Interactions*, no. 3, pp. 44–47, 2015.
- [4] M. Ludvigsen, M. H. Fogtman, and K. Grønbaek, "TacTowers: an interactive training equipment for elite athletes," *DIS 10 Proceedings of the 6th conference on Designing Interactive Systems*, pp. 412–415, 2010. [Online]. Available: <http://doi.acm.org.proxy.lib.sfu.ca/10.1145/1858171.1858250>
- [5] M. M. Jensen, M. K. Rasmussen, F. F. Mueller, and K. Grønbaek, "Keepin' it Real," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, no. April, pp. 2003–2012, 2015. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2702123.2702243>
- [6] J. Koekoek, H. van der Mars, J. van der Kamp, W. Walinga, and I. van Hilvoorde, "Aligning Digital Video Technology with Game Pedagogy in Physical Education," *Journal of Physical Education, Recreation & Dance*, vol. 89, no. 1, pp. 12–22, 2018. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/07303084.2017.1390504>
- [7] J. Matejka, T. Grossman, and G. Fitzmaurice, "Video Lens : Rapid Playback and Exploration of Large Video Collections and Associated Metadata," in *Uist*, 2014, pp. 541–550.
- [8] J. Vales-Alonso, D. Chaves-Dieguez, P. Lopez-Matencio, J. J. Alcaraz, F. J. Parrado-Garcia, and F. J. Gonzalez-Castano, "SAETA: A Smart Coaching Assistant for Professional Volleyball Training," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 8, pp. 1138–1150, 2015.
- [9] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3425–3434, 2017.
- [10] W. Pei, J. Wang, X. Xu, Z. Wu, and X. Du, "An embedded 6-axis sensor based recognition for tennis stroke," *2017 IEEE International Conference on Consumer Electronics, ICCE 2017*, pp. 55–58, 2017.
- [11] G. Bellusci, F. Dijkstra, and P. Slycke, "Xsens MTw : Miniature Wireless Inertial Motion Tracker for Highly Accurate 3D Kinematic Applications," *Xsens Technologies*, no. April, pp. 1–9, 2018.
- [12] X.-i. Technologies, "NG-IMU," 2019. [Online]. Available: <http://x-io.co.uk/ngimu/>
- [13] Y. Wang, Y. Zhao, R. H. Chan, and W. J. Li, "Volleyball Skill Assessment Using a Single Wearable Micro Inertial Measurement Unit at Wrist," *IEEE Access*, vol. 6, pp. 13 758–13 765, 2018.
- [14] J. Cancela, M. Pastorino, A. T. Tzallas, M. G. Tsipouras, G. Rigas, M. T. Arredondo, and D. I. Fotiadis, "Wearability assessment of a wearable system for Parkinson's disease remote monitoring based on a body area network of sensors," *Sensors (Switzerland)*, vol. 14, no. 9, pp. 17 235–17 255, 2014.
- [15] S. I. Ismail, E. Osman, N. Sulaiman, and R. Adnan, "Comparison between Marker-less Kinect-based and Conventional 2D Motion Analysis System on Vertical Jump Kinematic Properties Measured from Sagittal View," *Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS)*, vol. 392, no. 2007, pp. 11–17, 2016. [Online]. Available: <http://link.springer.com/10.1007/978-3-319-24560-7>
- [16] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs," *Computer Graphics Forum*, vol. 36, no. 2, pp. 349–360, 2017.
- [17] S. Harvey and C. Gittins, "Effects of integrating video-based feedback into a Teaching Games for Understanding soccer unit," *Agora para la educación física y el deporte*, vol. 16, no. 3, pp. 271–290, 2014.
- [18] D. Schulhaus, C. Zwick, H. Körger, E. Dorschky, R. Kirk, and B. M. Eskofier, "Inertial Sensor-Based Approach for Shot / Pass Classification During a Soccer Match," *Proc. 21st ACM KDD Workshop on Large-Scale Sports Analytics*, vol. 27, pp. 1–4, 2015. [Online]. Available: <https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2015/Schulhaus15-ISA.pdf>

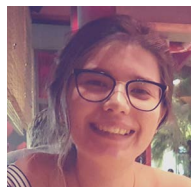
- [19] E. Mitchell, D. Monaghan, and N. E. O'Connor, "Classification of sporting activities using smartphone accelerometers," *Sensors (Switzerland)*, vol. 13, no. 4, pp. 5317–5337, 2013.
- [20] Weiping Pei, Jun Wang, Xubin Xu, Zhengwei Wu, and Xiaorong Du, "An embedded 6-axis sensor based recognition for tennis stroke," in *2017 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2017, pp. 55–58.
- [21] M. Kos, J. enko, D. Vlaj, and I. Kramberger, "Tennis stroke detection and classification using miniature wearable imu device," 05 2016.
- [22] P. Blank, J. Ho, D. Schuldhuis, and B. M. Eskofier, "Sensor-based stroke detection and stroke type classification in table tennis," in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ser. ISWC '15. New York, NY, USA: ACM, 2015, pp. 93–100. [Online]. Available: <http://doi.acm.org/10.1145/2802083.2802087>
- [23] L. Nguyen Ngu Nguyen, D. Rodriguez-Martn, A. Catal, C. Prez, A. Sam Monsos, and A. Cavallaro, "Basketball activity recognition using wearable inertial measurement units," 09 2015.
- [24] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, and Y. Liu, "Towards unsupervised physical activity recognition using smartphone accelerometers," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10701–10719, Apr 2017. [Online]. Available: <https://doi.org/10.1007/s11042-015-3188-y>
- [25] T. Kautz, thomas. kautz, and benjamin. groh, "Sensor fusion for multi-player activity recognition in game sports," 2015.
- [26] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity," *Neurocomput.*, vol. 181, no. C, pp. 108–115, Mar. 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2015.08.096>
- [27] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, "Activity recognition in beach volleyball using a deep convolutional neural network," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1678–1705, 2017.
- [28] L. P. Cuspinera, S. Uetsuji, F. Morales, and D. Roggen, "Beach volleyball serve type recognition," in *Proceedings of the 2016 ACM International Symposium on Wearable Computers*. ACM, 2016, pp. 44–45.
- [29] M. Kos, J. Ženko, D. Vlaj, and I. Kramberger, "Tennis stroke detection and classification using miniature wearable imu device," in *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2016, pp. 1–4.
- [30] Z. Zivkovic, F. van der Heijden, M. Petkovic, and W. Jonker, "Image segmentation and feature extraction for recognizing strokes in tennis game videos," in *Proc. of the ASCI*, 2001.
- [31] P. Jarit, "Dominant-hand to nondominant-hand grip-strength ratios of college baseball players," *Journal of Hand Therapy*, vol. 4, no. 3, pp. 123–126, 1991.
- [32] R. Velasco, *Apache Solr: For Starters*. CreateSpace Independent Publishing Platform, 2016.
- [33] L. P. Cuspinera, S. Uetsuji, F. J. O. Morales, and D. Roggen, "Beach volleyball serve type recognition," in *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ser. ISWC '16. New York, NY, USA: ACM, 2016, pp. 44–45. [Online]. Available: <http://doi.acm.org/10.1145/2971763.2971781>
- [34] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, "Activity recognition in beach volleyball using a Deep Convolutional Neural Network: Leveraging the potential of Deep Learning in sports," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1678–1705, 2017.



Fahim A. Salim is a Post-Doc researcher at Biomedical Signals and Systems Group, University of Twente. He is currently working on the Smart Sports Exercises project which utilizes IMU sensors and pressure sensitive in floor displays to offer tailored and interactive exercise activities in the context of volleyball training and analysis. Fahims research interest is to combine Multimodal Signal Processing and Human Media Interaction approaches in multi-disciplinary applications.



Fasih Haider is a Research Fellow in the Usher Institute, at the University of Edinburgh, UK. His areas of interest are Social Signal Processing and Artificial Intelligence. Before joining Usher Institute, he was a Research Engineer at the ADAPT Centre where he worked on methods of Social Signal Processing for video intelligence. He holds a PhD in Computer Science from Trinity College Dublin, Ireland. Currently, he is investigating the use of social signal processing and machine learning for monitoring cognitive health in the SAAM project.



Sena Busra Yengec Tasdemir is a PhD student at Electrical Computer Engineering Department, Abdullah Gul University, Turkey. Her research area involves Computer Vision and Pattern Recognition approaches. Currently, she is working on a project which aims to detect Breast Cancer from Digital Mammograms with the help of Computer Vision.



Vahid Naghashi is a PhD student at Computer Engineering Department, Bilkent University, Turkey. His areas of interest are deep learning, computer vision and time-series prediction. He worked on the project for 3D face reconstruction from 2D images and also image segmentation using evolutionary algorithms.



Izem Tengiz is a Bachelors student at Department of Biomedical Engineering, Izmir University of Economics, Turkey. She worked on a project which aimed to detect Bipolar Disorder from Magnetic Resonance Images of brains. Currently, she is working on her Senior Project which involves deep learning and image processing.



Kubra Cengiz is a PhD student and a research assistant at Faculty of Computer and Informatics Engineering, Istanbul Technical University, Turkey. Her areas of interest are Computer Vision, Medical Image Processing and Machine Learning. She currently works on designing machine-learning based models for predicting high-resolution medical data from low-resolution data.



Dees B.W. Postma is a post-doctoral researcher at the Human Media Interaction (HMI) group of the University of Twente. Dees currently works on the Smart Sports Exercises project in which he designs interactive digital-physical training exercises using an interactive LED-floor. In this project, Dees combines his research interests on perception and action, sports sciences and interaction technology to arrive at innovative training exercises for volleyball.

His work focuses on the interaction aspects of whole body interaction and steering behavior during play, looking at various contexts from stimulating movement to transforming social interactions, from sports to health, and from doing this for children to older adults.



Robby van Delden is an assistant professor at the Human Media Interaction (HMI) group of the University of Twente. His work focuses on the interaction aspects of whole body interaction and steering behavior during play, looking at various contexts from stimulating movement to transforming social interactions, from sports to health, and from doing this for children to older adults.



Dennis Reidsma is Assistant Professor at the Human Media Interaction group, Lecturer at Interaction Technology and Creative Technology, and Design-Lab Fellow, at the University of Twente. He investigates the transformative impact of interactive technology in play and learning in two areas. First, he leads a multidisciplinary team that works on human-robot and human-agent interaction in various scenarios of coaching and learning. Second, he pursues a research line on "play with impact" through various projects of playful interaction in smart environments, for entertainment, education, sports, and health & wellbeing applications. He has collaborated on the development of a number of interactive play platforms for childrens play, play for people with Profound Intellectual and Multiple Disabilities, play for gait rehabilitation, volleyball training, and other domains. Central theme in these projects is the potential for playful interactive technology to influence the social and physical behaviour and experience of the user.



Saturnino Luz is a Reader at the Usher Institute, University of Edinburgh's Medical School. He works in medical informatics, devising and applying machine learning, signal processing and natural language processing methods in the study of behaviour and communication in healthcare contexts. His main research interest is the computational modelling of behavioural and biological changes caused by neurodegenerative diseases, with focus on the analysis of vocal and linguistic signals in Alzheimers's disease.



Bert-Jan F. van Beijnum is associate professor at the faculty of Electrical Engineering, Mathematics and Computer Science. His research addresses smart technologies for remote monitoring, analysis and feedback technologies for patients with chronic conditions, support of lifestyle changes and sports. He is involved in projects on methods and technologies for monitoring stroke patients in daily life, development of rehabilitation devices for stroke, monitoring coaching and behavioural modelling of type 2 diabetes patients, qualitative assessment of rehabilitation after hip fractures, minimal sensing for motion capturing and running, modelling athlete behaviour for smart sports exercises.

ISBN 978-605-9788-33-5



Bilkent University
Department of Computer Engineering